The Dissertation Committee for Maria Esteva certifies that this is the approved version of the following dissertation:

THE *ALEPH* IN THE ARCHIVE: APPRAISAL AND PRESERVATION OF A NATURAL ELECTRONIC ARCHIVE

Committee:

Patricia K Galloway, Supervisor

David B Gracy II

Maytal Saar-Tsechansky

Victoria D Horwitz

Mary E Cunningham-Kruppa

Mary Lynn Rice-Lively

THE *ALEPH* IN THE ARCHIVE: APPRAISAL AND PRESERVATION OF A NATURAL ELECTRONIC ARCHIVE

by

María Esteva, MSIS

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2008

To my son Simón

Para mi hijo Simón

ACKNOWLEDGMENTS

My dissertation committee members deserve special appreciation. Patricia Galloway, taught me about the multiple layers and possibilities of digital records, and endured with me through a complex research project. David Gracy's passion and knowledge inspired me to pursue the field of Archival Studies, and it was when Ellen Cunningham-Kruppa mentioned that she was on board of the digital train that I decided to ride it too. Victoria Horwitz helped me look at my work and academics from refreshing perspectives. With intellectual generosity and assuredness, Maytal Saar-Tsechansky guided me into the territory of text mining. Mary Lynn Rice-Lively assisted my more basic needs of work and scholarships that allowed me to finish this degree.

I would not have been able to come to study at the University of Texas at Austin nor do this research if it was not for Fundación Aleph's support. While for confidentiality reasons I cannot name them, I deeply appreciate how staff members and information technologies consultants that worked there in various periods contributed to this project and trusted me with the product of their work: their archive. I am very proud of having been a part of an institution that I believe made a big and positive difference in Argentina. Also through my work there I met Carolyn Rose, whose teachings in preventive conservation have great influence in this dissertation.

Different people and groups have been very generous and helpful. The Seminar on the Acquisition of Latin American Library Materials (SALALM) granted me with the Marietta Daniels Presidential Scholarship throughout my doctoral studies. Others shared their talents in this dissertation. Shane Williams answered small and big systems administration questions, Leonardo Martínez contributed with precision preparing the

dark archive in Buenos Aires, and Paul Navratil from TACC worked on the animated visualization. Daniel Zeman lent me his language sorter, and Analía Sabán her beautiful art-work for my presentations. Very special thanks go to Hai Bi, programmer and math wizard, whose enthusiasm and curiosity enhanced my work in so many ways. Finally, my smart i312 “Information in Cyberspace” teaching team covered for me during a hard last semester in the doctoral program.

My son Simón and I met wonderful people in Austin that kept us company these years. My dear Kevin protected me and adopted Simón and through him we learned to love all about Texas. He, Bianca, Serge, and Nicolás are our family away from home. With Uri, Lisa, and Sue I shared work, classes, coffee, and confidences. Apen, Lucia, and Jackeline left Austin a while ago and I have missed them at the Colorado apartments as I will miss Rodrigo when we move. During these years I forged a doctoral degree and met new friends, but I lost birthday parties, christenings, gatherings, conversations, and sad and happy moments with my family and friends in Argentina. I miss them very much and value their generosity to let me go.

It never crossed my mind how much I was going to like Austin and this University in which I have lived, worked, and studied. For almost eight years I lived in the University of Texas at Austin Apartments for Families by the very green Colorado River. I have been very happy here; both Simon and I felt very safe and made unforgettable friends. For mythical reasons, the Colorado River connects me to other rivers that are important to me. I hope that this river takes us to other peaceful and bright places.

THE *ALEPH* IN THE ARCHIVE: APPRAISAL AND PRESERVATION OF A NATURAL ELECTRONIC ARCHIVE

Publication No._____

María Esteva, Ph.D.

The University of Texas at Austin, 2008

Supervisor: Patricia K Galloway

This research explores whether digital records created and used in environments without explicit record-keeping rules provide evidence of the organization that creates them and can be preserved in the long term. I studied the formation process of a digital archive that belonged to a philanthropic organization in Argentina. This archive originated in the late 1980s and was added to until 2005, a period during which, as information technologies were being massively adopted in the work-place, new problems were compounded by the nature and conditions of its electronic records and database systems. The study revealed knowledge about the information technologies and social practices used in the archive's development, providing an understanding of the path from its past to its present form and insights about how to preserve it. The attributes characterizing this archive led to developing the concept of *a natural electronic archive*.

To determine whether the records in the natural archive reflect the organization that created them I devised an inductive appraisal method that uses text mining, social network analysis, and visualization methods. I calculated the similarity between the text records created, gathered, and shared by them within frameworks of time and provenance as a measure of the strength of the relationships between staff members and the functions that they represented. Results of mining electronic text records belonging to 10 years of activities in the organization indicate that it is possible to observe changes in work-dynamics and roles in a way that goes beyond the typical organizational chart. The process and challenges involved in developing and validating the appraisal method are reported in this dissertation.

Studying the archive's formation process allowed gaps in the technical documentation to be filled and suggested a preservation strategy. The goal of the strategy is to preserve the structure and context in which the electronic records and databases were created and used, while moving them into a new and compatible technical environment to allow continuous access. From a practical perspective the strategy allows studying the effects of hardware and software migration on file formats and databases present in the digital archive. From a broader perspective it aims to provide a theoretical understanding of the relationship that exists between digital information creation and use and preservation strategies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTAL FILES

Use of Text Mining and Visualization to Infer Work Dynamics from Organizational Records. Quick Time movie.

PART I: ENTRY IN THE FIELD

Introduction

In his story *The Aleph*, writer Jorge Luis Borges describes a point in the cellar of an old house where the entire universe, past, present and future can be seen simultaneously, "without superposition and without transparencies."[1]I use the idea of *The Aleph* as a metaphor for the potential of identifying the various layers usually comprising digital archives and libraries, and to envision the research that needs to be undertaken to discover and preserve them.

This research answers key theoretical questions about the nature of electronic records, their evidential value, and their long-term preservation. Answering these questions thrust me into somewhat uncharted territory, where I had to not only develop ways of approaching the study of social interaction and work relations in institutions, but also to develop methodological tools of acquiring data that would allow for such a study. Indeed, because various aspects are inextricably connected, I am compelled to stress at this point—before further reading—that as a result of the need to integrate what until now were unconnected approaches, this dissertation will be of interest not only to archivists—namely a new way of studying a particular type of archive—but also contributes discussions and data to other fields of interest. I aim at: (i) understanding how people create, organize, and interact with digital archives; (ii) how these reflect their roles and activities providing evidence of their creators, and (iii) how digital archives can be preserved to reflect these activities over time. Specifically, I explore digital archives created and maintained in environments without explicit record-keeping rules and whose creation and development were never formally documented. This research is relevant

because it includes a large proportion of digital materials now extant that are at risk of being overlooked under a paradigm of regulated recordkeeping. I draw upon theoretical perspectives from Archival Science, Digital Humanities Research, Social Network Analysis, Information Retrieval, Preventive Conservation, and Material Culture and use a combination of qualitative and quantitative research methods including: case studies, interviews, formation process analysis, metadata extraction, text mining, analysis of social networks, and visualization.

I study the formation process of a digital archive created in the late 1980s and maintained until 2005, a period during which, as information technologies were being massively adopted in the work-place, new problems were compounded by the nature and conditions of its electronic records. The attributes characterizing the digital archive at hand led to developing the concept of *a natural electronic archive* that allowed transforming the entire archive into a unit of analysis. The study revealed knowledge about the technologies and social practices used in the archive's development; providing a path from its past to its present form and to insights about how to preserve it.

To determine whether the records in the natural archive reflect the organization that created them I devised an inductive appraisal method that uses text mining, social network analysis, and visualization. Measuring the similarity between text records created and gathered by staff members within frameworks of time and provenance, the strength of relationships between them and their functions was inferred. Results were validated against accounts of the staff members about whom they worked with, when and in what, through analysis of the contents of the records, and statistical distributions. After mining electronic text records belonging to 10 years of activities in the organization it is possible to observe the structure of the organization and changes in work-dynamics and

roles. Due to the uneven characteristics of the archive, the method presents challenges that will be reported.

Studying the archive's formation process allowed filling documentation gaps and suggested a preservation strategy. The goal of the strategy is to preserve the structure and context in which the electronic records and databases were created and used, while moving them into a new and compatible technical environment to allow continuous access. From a practical perspective the strategy allows studying the effects of hardware and software migration on file formats and the databases present in the digital archive. From a broader perspective it aims to provide a theoretical understanding of the relationship that exists between information creation and use and digital preservation strategies.

THE ROAD MAP

This dissertation is structured in five parts. The first four are interdependent articles and the fifth is the conclusion. Each part includes its correspondent conclusions and further research areas. In the first part I describe my entry in the field and how I found the topic of my dissertation in the midst of archiving the records of a private foundation in Buenos Aires. This experience constituted the preliminary observation of the institution's digital archive and led me to formulate the research plan in which I describe the tools and methods used to study it. By giving an account of the circumstances involved throughout the process of my work and my research, I clarify to readers my intentions and interventions and give them the opportunity to reflect about the extent to which my actions shaped the archive, my research, and my archiving work.

In the second section I study the archive's formation process, both paper and electronic, and give its electronic portion the name of natural electronic archive. The third section presents the digital appraisal method designed to find out if the electronic text records of the natural archive provide evidence of the organization. The fourth section describes the minimalist preservation strategy devised to stabilize the digital archive for the next ten years in which it will remain in retention period. In the last section I define the concept of the natural electronic archive and discuss the contributions and lessons learned.

An Archiving Project

In December of 2003 I suggested to the president of Aleph Foundation that the records of the institution be archived.[2] During the previous three years the foundation had been following a deconstruction plan that would conclude, in December of 2005, 20 years of philanthropic activities in Argentina. My proposal had the usual components of a typical archival plan, which included appraisal, inventory, arrangement, and preservation of the records. Its justification was based on the significance of the institution and the implications for Argentine society of losing its contents. The archiving project would be carried out in stages adjusted to my schedule as a doctoral student at the University of Texas at Austin and would conclude by the end of 2006.[3]

I had been associated with foundation since 1989 when I was awarded a grant to train in rare book conservation. Starting in 1993, I worked in several of the foundation's cultural heritage conservation projects in different museums and libraries. In 1997 I became conservation projects coordinator and full time employee. In 2000 I received

their support to pursue graduate studies in Information Science and Preservation Administration. Returning to archive the foundation's records, I resumed my former position as staff member. Knowing the institution and being known by the staff facilitated the archiving process and established an environment of trust. And yet, as the work progressed, I realized that my previous status as employee had not made me particularly knowledgeable about the institution's archive. Instead, I had been rather unconscious of my records creation and record keeping activities and those of others in the organization. This insight allowed me to approach the work with a fresh outlook and ultimately to discover the archive.

BRIEF DESCRIPTION OF THE ORGANIZATION [4]

Foundation Aleph was established in Argentina in 1985 to "conduct activities that help improve the living conditions of the community."[5] It was part of a broader philanthropic organization consisting of three sister foundations in different Latin American countries and an umbrella foundation that, within specific margins of autonomy retained by the sister foundations, controlled the flux of funds and oversaw their programs. By statute the foundation could not solicit money nor participate in fund-raising campaigns. It was legally bound to abide by local regulations for non-profit organizations. At the time it was created, its founders established a deadline to end the organization's operations, which occurred in 2005.[6]

Over a period of 20 years the Aleph foundation invested net 98 million dollars in fellowships and grants programs in the areas of Education and Sciences, Arts and Cultural Heritage, and Social Welfare and elicited investments doubling that amount in matching funds. At the time when it was set up, in the mid 1980s, this *modus operandi*

was something new on the local scene and became a model that would be followed by similar organizations. The foundation distributed the majority of its funds through lines of grants. It also supported and produced projects on its own initiative with the purpose of boosting innovative activities in its areas of interest. To help shape and to audit the programs, academics and experts in the areas of action were regularly consulted, and grant beneficiaries and projects were selected through a peer-review process. In general terms, the feasibility and application of the lines of support were determined by applying a set of common sense rules that emerged from assessments of the target fields vis-a-vis the foundation's vision and policies. However, flexibility was also allowed to accommodate exceptional situations and needs.

The foundation's organizational structure always remained small and hierarchical. It consisted of a board of directors, a president, an executive director, three area managers, and a financial manager. Most members of the board and the area managers were academics or experts in the foundation's fields of interest. For most of the foundation's existence, the executive director also managed the Education and Sciences Area until 2002, when an Education Manager was hired. The president and the executive director had secretaries and the area managers worked with a team of project coordinators and assistants. Additionally, during periods of intense activity, or for the production of specific projects, part-time assistants and project coordinators were hired to work for the project's duration. The administrative team, led by the financial manager, consisted of assistants and general services personnel. Through the years the foundation's staff remained relatively stable and never exceeded the number of 25 full time employees. Figure 1 below shows an organizational chart with the foundation's official structure and functional areas. The colors represent the different functional areas.

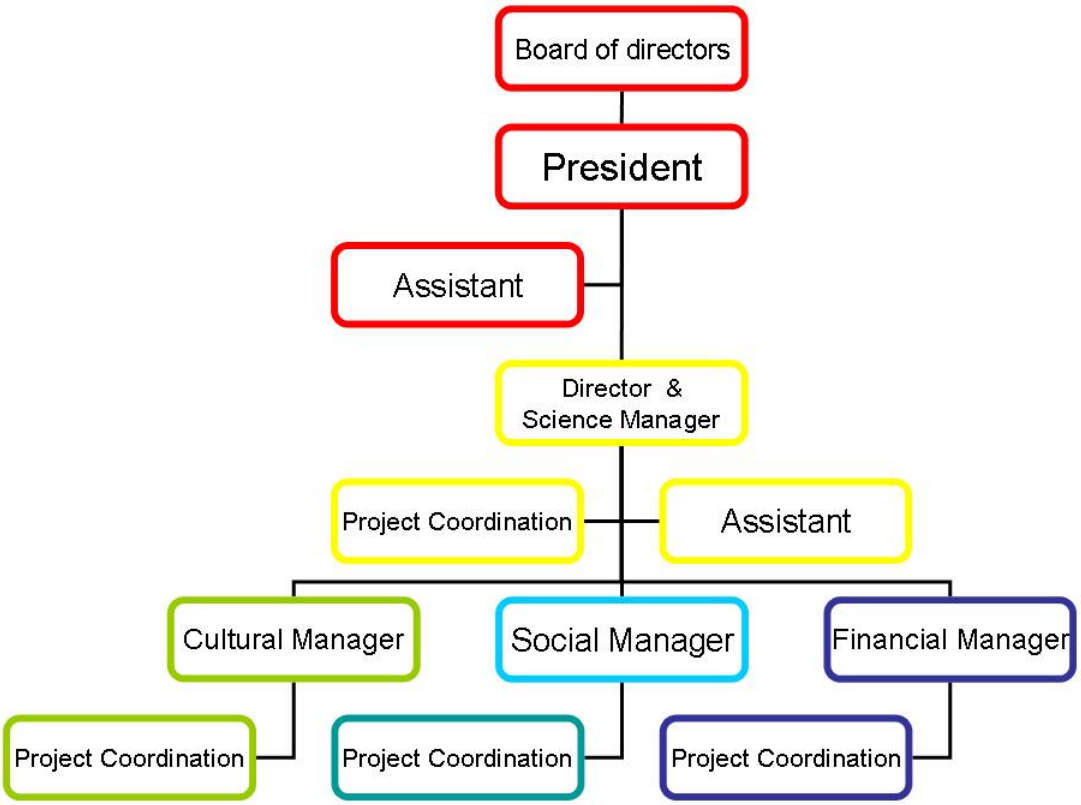


Figure 1: Aleph's organizational chart as of 2002.

In contrast with government funding agencies, Aleph was characterized by its public as diligent and punctual. Staff worked efficiently to accomplish their work with little overhead and in timely fashion and had a personalized communication with its beneficiaries. The institution was a pioneer in its early use of information technologies to manage grants and communicate with the public and with grant recipients.

A depiction of the foundation's programs and their impact is portrayed in the next section of an interview with the coordinator of a long-term Historic Photography Preservation Program supported by the foundation. His testimony is illustrative of the

intensity and the spirit with which projects were undertaken. Indirectly, it gives a hint of the amount of work that was involved.

> *When I say that Aleph was a miracle in our culture I mean that it was a providential event (although of course Providence had nothing to do with it), something unexpected and extraordinary on account of the absence of any preceding experience of similar nature, and whose activity and projection had a character of singular fertility. To make reference only to the preservation and dissemination of our cultural patrimony, an area that I know and about which I can speak with some authority, Aleph's activity set the standards for a social knowledge about scientific conservation principles for the country's cultural patrimony and for modern museum and archive management through: intense and extended seminars taught by specialists and professionals, especially European and American, of very highest level; complex projects that combined these seminars with practical activities of intervention in museums; numerous annual scholarships to develop knowledge abroad; diffusion of these interventions and of the patrimony, many times unknown or very little known to the public, through books of excellent and rigorous makeshift; and especially the creation and activities of a center for art restoration that was and still is unique in the country and in the region…*
>
> *The same considerations are applicable to the work done in the Argentine historic photographic conservation program: never in our country was a sustained and active interest displayed towards the photographic patrimony as was developed in that program. Over five years 400,000 dollars were invested in localizing and applying conservation measures in more than forty photographic archives, mostly public and also private, and training professionals with courses, seminars and in special cases with scholarships to train at the George Eastman House in Rochester. Likewise, during the twenty years that the foundation lasted, we worked to disseminate the Argentine photographic patrimony through books , which in all the cases, incorporated historical essays about the provenance of the prints and the activity of the photographers that took them. These books combine the display of our photographic history, research essays, and a concrete anthology of images. A sustained activity, applied to a segment of culture practically neglected for such a long period until then, was something absolutely unthinkable before Aleph and not likely to happen after Aleph, even though the fruits of the activity are valuable and appreciated.*
>
> *Another outcome of the Foundation's activities is that it attracted people of high intellectual level and, at the pace of common work, established personal and work relationships of deep and extensive repercussion to our culture. Would I have met _____if it was not for the Foundation? Absolutely not... and now we are embarked in a project that demands mutual trust and effort. I think that this is*

another proof of how extraordinary was the foundation's existence; what it left after it expired. As a summary I can only say that I feel very fortunate for having been part of Aleph's experience and I hope to extend it as far as I can.[7]

As a former beneficiary I can attest that receiving a grant from Aleph was competitive and prestigious, with the added bonus that payments were always on time and the process was run smoothly. Once the grant was bestowed, the beneficiary had to go through processes of renewal and give proof that they still deserved to be supported until the end of their program. I obtained a scholarship to train in rare book conservation at the Library of Congress, and after completing the study program I felt pride and accomplishment, and the responsibility to keep up with the foundation's standards.

Working at the foundation was a daily challenge: exciting, stressful, demanding, creative, and rewarding. With Aleph's support, both as a student and as a professional I had the rare privilege (which always comes with high doses of pain and effort) of turning ideas into concrete actions, first by submitting them to critical evaluators, then by producing them, and finally by reviewing the results with honesty and a desire to improve. The foundation made an intense positive imprint on me personally and professionally which is why I feel so strongly about preserving its legacy. [8]

GAINING CONTROL OVER THE PAPER RECORDS

In May of 2004 I went to Buenos Aires for three months to begin the archiving process. There, I learned that the foundation had to leave its building and move to a smaller office in the next eight months; we needed to move the paper records to outside storage. Transferring the records became the opportunity to gain intellectual and physical control over them. I started by conducting a macro-appraisal of the institution and of the paper records.[9] The procedure followed was to analyze the files found in the office spaces

that corresponded to each of the foundation's organizational units and to identify records series and the function or sub-function to which they belonged.

Each organizational unit had file cabinets with area specific records; organized and named according to the criteria imposed by each area's project coordinators and assistants. Certain administrative files, such as board meeting minutes or action plans and annual budgets were maintained in chronological order in their respective area of provenance. The bulk of the records consisted of project files. While grants were active, their project files were maintained in the correspondent area offices; after the projects' completion, the folders were stored in compact shelving by year and project number in the building's basement, which was referred to as the archive. This project file archive constituted the only centralized file system in the organization that everybody used and understood.[10]

The results of my analysis became the outline that we used later to arrange and inventory the records. The first level of arrangement—group level—corresponded to the institution's organizational structure. The record series titles emerged from the way in which records had been grouped and/or named by staff members. Example series titles are: correspondence, board meetings, luncheons, events, publications, trips, marine biology, museums, music, building, personnel, etc. For purposes of inventory control, I devised a numbering scheme that corresponded to the foundation's hierarchical structure and to the records series so that the physical arrangement of the records would correspond to their original order and provenance.

With the purpose of finding off-site storage space for the records, two records management companies were invited to submit proposals and bids.[11] One of the issues that emerged during our preliminary conversations with them was their lack of

understanding of the difference between an archiving project and their standard records management services. Our project goal was to create a comprehensive guide to the body of records that would allow understanding its structure and components. Instead, the companies suggested storing the files in boxes with succinct labels and controlled by a bar code which would be useless for accessing the records by provenance in the future. We were also concerned with the long-term preservation of the paper materials and wanted to re-house the items with new archival quality enclosures.

An alternative was offered by the foundation's education manager, who suggested that we could conduct the re-housing and inventory of the records ourselves. The staff members—already reduced in number—knew their records well, and since their former activities at the foundation where winding down, the work could be accommodated without much additional work. This solution would require training existing staff in archival principles, records inventory, arrangement and re-housing, and coordinating our records' arrangement and numbering system with the records management company's control system. We decided that this was the better approach, and I wrote a document for the bidders clarifying their duties and our responsibilities so they could prepare a budget according to our needs. Within a month, we selected a company that accommodated to our conditions and budget. In its storage facility the company provides security measures such as fire, collapse, and theft prevention, but like all such facilities in Argentina it does not provide air conditioning for bulk paper cartons due to the costs that it entails.[12] For housing and moving the records, the records management company would supply the cartons and we purchased archival quality filing supplies to re-house our records. A workflow stating the number of cartons and the frequency with which we would receive and fill them, and the company would pick them up was established.

The staff training sessions were held in the foundation's conference room. After two presentations on archival principles and processing to introduce the staff members to the broader and practical aspects of the work to come, we proceeded to complete hands on inventory, arrangement, and re-housing according to the results of the macro appraisal. As the staff analyzed the records pertaining to their functions, we clarified questions and refined the process. As a group we reviewed and corrected the structure and functions that I had identified, decided what type of materials should be considered non-records, and agreed on the meaning of confidentiality and sensitive information in light of the records at hand. While records were being processed we also made modifications to the fields in the inventory spreadsheet that I had prepared. In it, staff members had to enter information at the series level concerning: provenance, date ranges, subject treated, confidentiality, and completeness. Once the records were re-housed and located in their cartons, the staff noted folder and box numbers as well as the identification assigned by the records management company. At the same time, I was writing a manual with instructions for inventorying, arranging, and housing of records that would serve as a reference to staff members and document the process.

Reviewing the legal accounting and tax books[13] I noticed that portions of them were already illegible or in the process of becoming so due to ink fading.[14] To avoid further loss of information and to leave a record of the missing data, all the books were microfilmed. The preservation masters were deposited in an environmentally controlled storage room in a microfilm archive and will not be made accessible until—and if—the foundation executors authorize it. We also found the ink fading problem in most faxes enclosed in the files and decided to make preservation photocopies of those still legible.

During the training sessions, memories and anecdotes elicited by the records were seasoned with the foundation's traditional three o' clock coffee and pastries break. One of us read the list of attendees and the menu served at one the institution's exclusive luncheons; other commented on a letter full of impressions sent by an artist about her training experience abroad. A young beneficiary from a music training program had pasted to her progress report all the bus tickets used to commute from her home town to the main provincial city to take piano lessons. This was accompanied by a letter of her professor praising her efforts and commenting on her improvement. The number of projects that were approved during the monthly board meetings surprised us as if we were only then becoming aware of the amount of work that we had accomplished through the years. At the time I thought that the sessions had a healing effect; at the end of foundation's activities, they allowed staff members from all the areas to start the process of closure.

LEGAL CONSIDERATIONS

Confidentiality, copyright, laws, regulations, and internal policies were addressed with the purpose of exploring donation prospects and records retention requirements. I was concerned about the feasibility of donating the project files that contained records of the entire grant process, from selection to progress evaluation and project completion. These files contained grant information that the foundation had committed to maintain confidential and personal information that had to be considered vis-à-vis Argentina's "Personal Data Protection Act." Also, I wanted to recommend what to do with the audiovisual collection that mostly consists of the works sent by the beneficiaries for progress evaluation or as final products resulting from their awards.[15] As a first step I

decided to gather information from different sources which, along with the results of the records' analysis, would allow me to make recommendations.

To learn about the criteria used to determine access restriction for foundation archives, particularly concerning project files, I wrote an email with questions to the Rockefeller Archive Center. This institution collects papers of philanthropic organizations as well as its own Rockefeller Foundation records and has a leadership role in helping foundations establish records management and archival programs. The Center's Assistant Director sent me a clear description of their access policies.[16] While specifics are negotiated with each donor, as a general practice to protect the privacy of those involved in the awards decision making process, public access to grant files is restricted for twenty years after the grant's completion. Officers' diaries, containing notes about meetings and accounts of trips abroad, are also closed for twenty years after they are written or until their authors are deceased. Fellowship files are permanently closed to protect the confidentiality of the transactions between the fellows and the agency; however a synopsis about the recipient, the amount of the award, the object of study, and excerpts from interviews or correspondence are made available for study.

I also asked the foundation's lawyers to give their opinion about prospects for donation to a permanent archive based on the analysis of examples of typical project files and considering the foundation's policies and grant conditions as well as local legal regulations. Requirements pertaining to a prospective donation were analyzed in light of the Argentine "Personal Data Protection Act" Law 25.326. In its first article, the law establishes that the purpose of the Act is,

> *..the comprehensive protection of personal data included in files, records, databases or other data processing technical means – whether public or private – used for reporting purposes, in order to guarantee the right of individuals to their honor and privacy as well as access to information recorded thereupon, in accordance with the third paragraph of Article 43 of the National Constitution.* [17]

After carefully analyzing this law, the lawyers considered that since the applicants agreed to give their information to the institution for purposes of managing grants, and the institution was not a legal person holding records or databases—as are the project files—with the function of producing reports for the general public or selling the information to other companies, the project files did not fall under the Act's requirements.

Nevertheless, their final recommendation was that the project files could not be made public. This was based on the foundation's confidentiality statement, published in the foundation's annual reports and on its web-site The statement stipulates that information about applicants and beneficiaries would only be used internally and for the purposes of making decisions and evaluating progress and would not be divulged to third parties; any exception to the rules would require the express consent of those involved. Therefore, one way of considering donating the project files would be to ask beneficiaries for their permission. Another option—similar to what the Rockefeller Archives Center does with its own project files—would be to create a synthesis of those files for the purposes of research. In fact, this synthesized version already exists in the foundation's grant tracking database.

The case of the audiovisual collection was fairly straightforward. Most of the tapes, CDs, DVDs and such, entered the collection as background material for applications, progress evaluations, or to be used in training programs and meetings. Kept separately from the project file for housing purposes, they were controlled through a

database in which their provenance (project number) was recorded. In most cases, the purpose for which they had been submitted, that is to be evaluated for a grant, fell within the scope of the confidentiality statement, and therefore the materials could not be donated. In terms of completion, the status of these materials ranged from complete and in progress to unfinished, in other words from copyrighted to un-protected, so donating them was not an option.

I also asked the foundation's lawyers to determine the period of time that the foundation had to legally retain its records after the end of activities. Because in Argentina there are no regulations specific to records of civil associations or foundations, they considered that the stipulations of the Code of Commerce whose articles 44 and 67 establish a retention period of ten years for accountancy and tax documentation of commercial companies applied to this case.[18]

Finally, I looked up what the Law 15.930, General Archive of the Nation (AGN) of 1961, says about records of private organizations and about records retention schedules. The General Archive and other governmental archives (provincial and municipal) accept records of private organizations on a "case by case basis." [19] The law quotes the Code of Commerce stating that those institutions that are regulated by this code have to retain their records for the required period of time before they can be transferred to an archive.[20] The General Inspectorate of Justice—an agency that controls the registry of public commerce and civil associations and foundations overseen by the Ministry of Justice and Human Rights—is responsible for identifying possible archival cases and of informing the appropriate governmental archival institution. To the extent that I could inquire, this clause is not enforced.[21] If acquired by a public archive, private records will be made available to the public after fifty years from the time which the

General Inspectorate of Justice designates as that of dissolution or extinction of the creating organization.[22] An annex to the law issued in 1981 defines records retention schedules for personnel records and for registry control records.[23] Except for the personnel records, the rest of the Argentine government document types and the government record-keeping system are difficult to match with those of philanthropic organizations and with the concept of record series.[24] Still, records proper to foundations such as board meeting minutes and project files, can be paralleled to resolutions, acts, and ordinances that in the government filing system—and differently from a series based filing system—are normally included within a case file and considered worthy of permanent retention. Other types of records that foundations and commercial institutions alike are obliged to maintain by law are: accounting records, board minutes, and tax books.

Overall, my reading of all the local recommendations and sources was that none have deeply considered the long term social implications of not retaining the records of organizations such as Aleph or others. It seems that as a society we do not have a culture of documenting ourselves or a collective notion of the accountability and historical role of societal archives and therefore recommendations about what to do are always partial. This leaves the owners of archives to decide what their place in history will be. After analyzing the information gathered, my recommendation was to retain all records for a period of ten years after the official cessation of activities as specified by the lawyers in accordance with the Code of Commerce. Beyond that, I suggested that the foundation should explore alternative options to donate the archive to a research institution such as a university archive, under conditions that would assure that the confidentiality clause is respected. This could mean asking former beneficiaries to authorize the donation of their

project files, synthesizing the information from the project series (already synthesized in the grant tracking database) or donating everything except the project files. I also presented the reservation of whether the confidentiality clause applied beyond the death of the parties involved.

MIXED FORMATS

That winter of 2005 the priority was to process the paper files to help empty the building, but I also spent a portion of my time analyzing the electronic records and exploring possible preservation strategies. When I arrived to work on the archiving project, the systems administrator set up a computer station for me with a password to log on to the network. I realized then how much I had forgotten about the way in which I used the network during my tenure as a staff member.[25] In hindsight, I can say that at that time I did my work without being totally conscious of my electronic record-keeping practices or the difference between storing my records on a shared directory instead of on my computer's hard-drive. Coming back with the perspective of a trained digital archivist, I re-discovered the use of the shared directory located on the networked server in which each staff member had a virtual folder with his or her initials to store the records that they created and gathered throughout their daily work.[26] The oldest file in that directory dated from 1991, and I even came across my own folder in the directory.

In various offices I also found 5.25" and 3.5" floppy disks for DOS and Windows operating systems containing programs, data, and texts. With the help of the systems administrator, who installed old drivers and software he had found (and had not thrown away), we used the DOS platform present in the Windows 98 operating system to access Professional Write and Lotus 1-2-3 files. From their original environments we migrated a

sample of Professional Write files to MS Word version 3.0 for DOS and then to Word 98 for Windows, and Lotus 1-2-3 files to Excel 98 for Windows. The exercise made me consider the possibility of recovering information in electronic format that I assumed was hidden in the many poorly labeled or non-labeled diskettes and disks that we found here and there in the offices.[27]

I also paid a visit to the information technologies (IT) consultant who had worked for the foundation since 1997. His consulting company designed and maintained the third iteration of the grant tracking and financial systems, and I wanted to understand the technology behind it and the code's ownership. He told me that the databases were created with Clarion 1.0 [28] for Windows for which they as developers paid the license, and that there were no contracts, agreements, or specifications regarding the code's ownership. In fact, I was the first person to bring that up, as it had never occurred to him or to the foundation to discuss that issue. During that meeting we talked about alternatives to migrate the data to an open source database system and also about alternatives to preserve the electronic records stored on the shared directory. His suggestion was that the server could be kept "as is," with all its contents and unplugged. I argued that this was an option that would not allow accessing the records if needed nor moving them into the future.[29] And yet, with modifications that I detail in the preservation chapter, this first suggestion was the platform for what became the minimalist preservation strategy.

After my brief re-encounter with the electronic records I thought that they could constitute my dissertation's case study (although at that point I did not know exactly what I was going to do with them.)[30] I asked the foundation authorities if I could use their electronic archive as my case study, and they generously said yes. To start a preliminary

exploration I copied, with its creator's permission, the sub-directory containing the files of the staff member who had worked for the longest period in the organization. Upon returning to Austin, my plan was to devise the electronic records archiving plan and my dissertation proposal based on a study of this body of files.

From September 2004 to May 2005 I supervised the transfer of the paper archives to the records management company through email and chat with the staff member in charge of coordinating the activities. The work was carried out smoothly and efficiently, and neat boxes with preservation file folders perfectly aligned and labeled were transferred to outside storage on a weekly basis. Pictures of the staff members with masks and gloves conducting archival processing activities were included in the foundation's last Annual Report issued in December of 2004. During that period I also started devising the electronic records archiving plan.

When I presented the original archiving proposal to the foundation I envisioned that all records, regardless of their format, would be arranged according to the structure that would emerge from the macro-appraisal analysis and that the electronic records would fit the same structural and functional categories as the paper ones.[31] However, after analyzing the structure of the virtual folder that I brought to study in Austin, I concluded that it would not be possible to use the same arrangement for both the paper and the electronic records. In the paper system, each organizational unit in the foundation constituted a sort of centralized record-keeping sub-system in which secretaries and project assistants filed their records according to functions and emerging projects. On the shared directory, on the other hand, each employee created his or her own record-keeping sub-system within which he or she grouped, named, and maintained the records

according to their own functions and projects and differently from other members of the same unit.[32]

To preserve this diversity, I explored quite deeply the possibility of installing DSpace [33] on the foundation's intranet and having the remaining employees upload their own records and those that belonged to other departed employees. For purposes of maintaining the records until further decisions about the future of the archive were made, we would find a secure web host for the DSpace implementation. Similarly to the procedure by which only authorized individuals could request paper records from the records management company, this DSpace repository would be accessible only to certain users. I wrote a detailed manual with instructions for the employees to conduct a preliminary analysis of their electronic records in preparation for the DSpace implementation. The manual described step by step instructions on how to identify file properties (dates and formats), determine levels of aggregation and topics, establish confidential record groups, and record the information on a spreadsheet in which I had previously established the fields. I planned to use this data to re-create the shared directory and subdirectories structure in DSpace by transferring them to communities and collections and granting the proper authorizations. This plan called for staff members to start the electronic records analysis in February of 2005 and their upload by mid June after I trained them in the use of DSpace. As records were uploaded, staff members would test their renderability and record the results that could inform a preservation decision. I presented the electronic records archiving plan to the foundation along with a schedule and the manual of instructions. Their response was that I should come to Buenos Aires to further discuss the proposal.

In the meantime I was moving forward with my research plan, in which I tried to combine the practical needs of the archiving project and the theoretical ones of my dissertation. I prepared a proposal for the University of Texas at Austin Institutional Review Board (IRB) to interview the foundation's staff members and the IT consultants who were working or had worked in the institution on the creation and development of the electronic portion of the archive and the technologies used over the years. Also, I wanted to learn about the record-keeping and records-creation practices, formal and informal, used individually and collectively. While at that point I did not have a perfectly delineated research plan, I decided to go ahead with the interviews because I was not sure that I would have other opportunities of traveling to Buenos Aires in the future.[34]

That Spring Semester I was taking a Data Mining class in the Department of Information and Risk Management in which we were briefly introduced to the concept and uses of text mining for competitive intelligence and knowledge management purposes.[35] I thought that I could mine the foundation's electronic text records although at that point I had no idea for what purposes. I spoke with my professor, and she agreed to guide me in an Independent Study in the Fall to learn more about the tools and techniques and to explore research venues for the electronic texts. With the purpose of bringing the electronic archive back to Austin for my dissertation, I asked the system administrator at the foundation how much storage space was occupied by the electronic records and purchased a 20 gigabyte portable hard-drive, which was twice the size that he had told me I needed.[36]

PAPER VERSUS ELECTRONIC RECORDS: FIRST AND SECOND RESEARCH QUESTIONS

By June of 2005 the foundation had moved to a smaller office and the majority of the paper records were in outside storage. Most staff members had left, and few remained to close the foundation's administrative affairs including project files belonging to the last call for grants. As soon as I arrived in Buenos Aires I met with the foundation's authorities to discuss my proposal, and they asked me the following question: Why was it necessary to keep the electronic records when all legal accounting and tax books and paper records are preserved?

The issue introduced a dichotomy that I had not considered: the competition between paper and electronic records. Which ones have legal value? Which ones are original? Is it worth it to keep both? Implicit in these questions is the issue of whether electronic records provide evidence vis-a-vis their signed and sealed paper versions. At that time I suspected that the electronic records could offer a different perspective on the foundation: that of the activities undertaken by the individual employees. But this was only a speculation on my part. I knew that I wanted to preserve the electronic records, but for this I needed to explore them to find out what it is that they could tell about the organization and to demonstrate whether they were an integral part of the archive. I would later turn these issues into two dissertation questions.

Of consideration also was the disarray in which the electronic records had been maintained and the efforts needed to organize them which, as I had proposed, implied a considerable investment. I paralleled the foundation's concerns with those of archivists and records managers who, taken aback by the state of chaos they encountered in examining servers and hard-disks, either consciously avoided dealing with them,[37] considered them "redundant, outdated, and trivial (ROT)," [38] or proposed approaching

them one by one, as they would a bundle of unorganized paper items.[39] Indeed, the tone of the discourse in "Review of the English Literature in Appraisal of Electronic Records" by Terry Eastwood reveals the anxiety experienced by archivists as they started to explore these systems composed of hidden databases and volatile interfaces, full of technical dependencies, in need of constant tweaking and migration, recording and erasing transactions as needed by ever-changing functions or due to incomprehensible errors.[40]

The prospect of disposing of these records made me reflect on the complexities and limitations of private institutions when they are simultaneously faced with the task of closing down and deciding the fate not merely of their archives but perhaps of their place in history. These concerns are further compounded by the characteristics of electronic records. Indeed, technological obsolescence gets in the way of the mourning period needed to achieve administrative and emotional distance from the affairs involved in closing an organization. Traditionally, this is resolved through retention periods during which archives remain closed. This was true for the paper portion of the archive, which was prepared to remain dormant in secure storage for the next ten years. But a similar interval conflicted with the promptness needed to take preservation action for the electronic records, and there were concerns about costs and human resources needed to maintain them accessible during the retention period at a minimum.

PRESERVATION ATTEMPTS: THIRD RESEARCH QUESTION

The option of storing the records in a DSpace repository was dismissed. Organizing and making the electronic records available within a repository system was not a priority for the foundation authorities, who were able to retrieve paper ones from

the records management company. The same reluctance applied to the idea of migrating the data from the grant tracking and financial systems to an open source database. I had another conversation with the foundation's IT consultant, who insisted on his original idea of keeping the records and systems in their current server and unplugged. He further suggested that the server could remain under the custody of the non-profit academic telecommunications network agency that provided Internet and email services to the foundation and with whom there was a longstanding relationship.[41] While this agency did not specialize in data storage, it had the expertise and the trust of the foundation to become the records' custodian for the next 10 years. I immediately arranged a meeting with the agency's coordinators during which we discussed custody options.

A preliminary idea was to keep the server running and the records and databases functional within their original software so they could be accessed for legal or administrative purposes. Backup copies of the server contents would be made and distributed to different storage locations for data security purposes. To prevent and address foreseeable mechanical failure and technology obsolescence, one of the agency's systems administrators suggested purchasing a new server and making budget provisions to buy replacement parts and to update the equipment; they would not charge for maintaining the server. My immediate duty was to prepare a document specifying the steps needed to complete an archival transfer of the records to the new server and the tasks required to maintain and control it for the next 10 years. I also had to build an inventory-guide that would allow accessing the records. Purchasing a new server implied a migration to newer software and hardware for which I had to determine how files and systems would function in their compatible but newer environment.

The second set of research problems was shaping. In order to consider archival trustworthiness and accountability, I had to create a protocol specifying operations to ensure maintenance of the original file properties during transfer to the updated server. Storage specifications had to assure maintenance of the integrity of the records and databases to protect their authenticity while guaranteeing that they remain accessible during the next 10 years. Beyond this initial period, I had to consider the possibility that all or parts of these records and systems could be retained for the long term. Underlying the proposal, the strategy of preserving the bitstreams guaranteed the possibility of migrating or emulating them in the future. All of these considerations would be later articulated as another research question for my dissertation.

FINAL PROPOSAL

In my new presentation to the foundation I emphasized keeping all the records, paper and electronic, for accountability purposes. In support of this recommendation I articulated a reason for the perceived dichotomy of paper vs. electronic and explained why the archive should be considered as a union of both. Finally, I laid out a stabilization strategy for the electronic records and systems that would not imply high costs or complex technical maintenance. I argued that while in our society the use of computers to record transactions and to create records was becoming dominant, we were still in a transitional period in which traditional conceptions and uses of records and record-keeping practices persisted along with new forms and uses supported by emerging technologies. Along these lines, the institution had used and maintained both systems since the early 1990s, and furthermore towards the end of its activities it started the process of announcing and awarding grants electronically. Still, at the time of closing, the

traditional perspective persisted, and instead of being seen as combined and complementary, the two systems were perceived as overlapping and repetitive.

I explained that certain electronic record types created by the institution constituted unique expressions whose content, form, and function could not be quite reproduced in paper format. This was the case with the institution's web-site, emails, and data within the financial and grants tracking databases. Over the years, these different types of records had sustained key operations through which investments, transactions, and decisions had been made. In both paper and electronic cases, creators had decided what to keep, what to duplicate, and what to discard as the events were taking place. Both electronic and paper records had been kept until the closure of the institution, and both constituted the archive. I did not see justification for the destruction of one or the other—or one over the other—"after the fact."

Moreover, I argued that destroying the electronic records could interrupt the ability of the institution to remain accountable in the eyes of society in the near future and prevent its archive from providing evidence in the long term. I pointed out that increasingly, electronic records and systems are subject to legal and administrative compliance. At that point, we had insufficient knowledge about the nature and content of the institution's electronic records to decide on their destruction. What would society say if they found out that records had been discarded? What if the foundation decided to donate the archive? Briefly, I also introduced the idea of research value, explaining that this could be the first electronic archive preserved in Argentina.

I specified that all the contents of the server should be stored in a new server and backed up on other media under the custody of the telecommunications agency at a minimum for the next ten years, using the same directory structure in which they had

been kept by their creators. For accountability purposes, I would generate a registry showing the directory and sub-directories structure, their contents, the records' format, and their creation and last modification dates. Based on the registry I would create a descriptive guide to help access the databases and the records. I would also specify archival standards for transfer and maintenance of records and for event reporting that the custodians would have to follow. File rendering software and an audit and control program would be installed on the server, which would be equipped with a dedicated monitor. The server would not be connected to the Internet.

The foundation authorities agreed to the new proposal and allowed me to use the copy of the electronic records for my dissertation providing that I did not disclose content. For reasons that are beyond the scope of this study, the electronic records did not remain in the custody of the telecommunications agency. Instead, it was decided that they would be kept by the last president of the organization. Still, the knowledge that resulted from the meetings in which we discussed the transfer plans and from the documents that I developed became the bases of the preservation strategy. I suspected that a big part of the reason why the foundation decided to maintain the electronic records was their support for my dissertation project. But I also thought that it was not easy to decide on discarding records accumulated over so many years and perhaps also under the influence of my characterizing them as an archive.

IN THE FIELD

While writing the proposal I was conducting the interviews with former and current staff members and IT consultants. Fortunately all the former employees that I planned to speak with lived in Buenos Aires, which allowed me to roam through many of

the city's beautiful neighborhoods during that sunny winter. The interview asked about the ways in which staff members created, stored, and circulated records; the technologies that they used; the types of records that they generated; with whom they worked. Their answers described work-processes, decisions, and actions. Writing a project, circulating a memo, sending or receiving an email, entering data or updating the databases, signing the letters offering grants and receiving the answer of acceptance, and filing the approved project constituted the actions through which plans, goals, and conversations were substantiated. We also talked about how decisions about information technologies unfolded in the organization and the different hardware and software used over the years. I noticed that with the exception of those particularly involved in the IT decision making process and its implementation (and just as had been the case for me), many interviewees did not know much about the tools used over the last twenty years, nor could they remember the names, dates, and versions of hardware and software with precision. In other words, they knew how to use what they needed for their work, but in general were neither savvy nor curious about IT.

I also resumed the analysis of the electronic records, this time with the intent of doing it systematically. The networked server, to which I had logged in so many times and whose structure I had navigated to access the staff members' directories, included other directories that I had never looked at. To make sure that I was considering all the foundation's records, I decided to survey all the contents of the server starting from the first directory level. During a preliminary browse I observed that here and there, in unlikely places within nested sub-directories, amongst application files, or loose under the first level directory, as if they had been misplaced, there were image and text files belonging to projects. With the help of the systems administrator we identified all sorts of

digital objects such as: email mailboxes and email backup scripts, old and new file management applications, older versions of the foundation's grant tracking and financial databases and their latest iterations, folders with images, printer drivers, and other program files for which at that point we could not determine origin or function.

The directory and sub-directories that specifically contained the staff members' records showed distinct characteristics. The majority of the files were texts, with a smaller proportion of spreadsheets, images, and presentation files. Each employee directory showed a unique structure and naming convention for folders and files. Within each folder, individual preferences ruled how records were named, used, organized, kept, and discarded. Complete and incomplete documents, versions, and personal records were ubiquitous, and across folders, similar fragments of text constituted the core of many documents.

Since the interviews were carried out in parallel with the survey of the server, I was able to clarify with the interviewees my findings on the server and vice versa. Looking at the server contents and listening to their stories about record creation, record-keeping, and the systems used over time I started to think about the networked server as a work-place. Over the years, all the staff members in different capacities and making different decisions had a part in its creation. Without much afterthought and in a natural way, records and applications were dropped, kept, and deleted, resulting in strata of different types of digital cultural material that, by the end of the institution's existence, resulted in a kind of digital version of an archeological site, only the most recent layers of which remained functional.[42] This notion reinforced the decision to keep the contents of the server intact, and the decision extended beyond the staff members' directory to keep the rest of the directories. I realized then that if I had pursued my original idea of

installing a DSpace implementation to upload and maintain the foundation's electronic records, I would not have had the opportunity of exploring this context. The idea of the server as a digital archeological site and the digital objects in it as remains from a work-environment evolved into the concept of the "natural electronic archive" that I describe later in this dissertation.

Up to this point, my work was shaped by emerging circumstances, by the demands and ideas of those with whom I worked, and by serendipity. The next steps, completed in 2006 and 2007, were part of my dissertation research plan. They included two more visits to Buenos Aires during which I completed the digital preservation strategy and met with some interviewees to resolve doubts and confirm results. The research plan is outlined in the next section and the results of these visits in the preservation chapter.

Research Plan

To explore the natural electronic archive I turned the networked server itself into an object of analysis, and the practical problems that I encountered in the process of archiving it became the following research questions:

1) What is Aleph's archive and what is a natural electronic archive?
2) What do the records of the natural archive tell us about the organization that created them? or alternatively What type of evidence of the organization is provided by the records of the natural archive?
3) How should this natural archive be preserved?

Around these questions I organized a plan that included three distinct parts. In the first one I researched and described the social and technical processes by which electronic records and systems had been created and used over twenty years to form what I define as a natural electronic archive. The second part involved creating and testing a digital appraisal method to determine what evidence about the creating organization was embedded in electronic text records belonging to archives thus characterized. For the third part I devised and implemented a preservation strategy for the natural archive so that the integrity and authenticity of the records and objects that accumulated over twenty years in the successive networked servers used in the institution could be preserved and accessed for the long term.

I researched these issues separately yet simultaneously and conducted the research as I was doing the archiving work. The digital archiving process focused on the networked server, and as I transitioned from work to research, the server—conceived as a work-space that evolved into a digital archeological site—was my unit of analysis. The answer to the first question is constructed throughout the dissertation. The narrative of the archive's formation process describes the origins and evolution of the systems and records found on the networked server and describes their characteristics. This encompasses the decisions that led to their existence and how they changed as a consequence of new technological developments, the different ways in which they were used over time and their interaction with the paper record-keeping system in existence. But in order to define the natural archive as a concept, I needed to incorporate the results of the appraisal exercise and the steps that led to designing and activating a preservation strategy for an archive of these characteristics. Therefore, the definition of a natural archive becomes the dissertation's conclusion. In the next section I describe the methods

and tools that I used for exploring the networked server and explain how the different sets of data that I gathered were organized and used to answer the research questions.

DISCOVERING THE ARCHIVE

The archive viewed reflexively

To manage my intervention as a researcher, records creator, and archivist in this case, I used an approach borrowed from reflexive archaeology. The goal of reflexive archaeology is to contextualize the research methods themselves. That is, to study a site —in this case the contents of the networked server—from the perspectives of the different contexts involved in its creation and development, including what occurs during the time in which the study is being conducted, but also if and how the methods of analysis influence the site's condition.[43]

Among the contexts involved in the study of Aleph's archive are: the staff members' views about their work practices and the decisions they made in relation to records-making and keeping; the status of information technologies in Argentina both in terms of human and technical resources; the privacy concerns proper to dealing with the archive of a private institution that includes personal information; the legal requirements that the foundation had to follow to close the institution; and the considerations made about the archive's future. Given that I was conducting my research and archiving the foundation's records at the same time, a major element to account for is if and how the methods used to preserve the archive shaped it. That is, I needed to know whether the authenticity, integrity, and completeness of the archive were maintained throughout the preservation and appraisal processes, all topics that are discussed in the preservation part of this dissertation. Finally, and given that this archive was created unintentionally, I had

to find out to what extent the staff members' idea of what constitutes the foundation's archive was influenced by my insistence on preserving the electronic portion.

From 2004 when we started archiving the paper records until the end of 2006, the staff members who were still working in the organization were trained to organize and describe their own paper records. Hearing about the discussions about what to do with the electronic portion, they were doubtless becoming more aware of the archival potential of these materials. During the auto archiving work the organization was still functioning and people were still creating records. While the process was informed by the idea of keeping it all, [44] inevitably, throughout the two years of archival processing, staff members' became more aware about their record-making and keeping practices and the archive was not created as "naturally" as it was before.[45]

Research methods and tools

To study the server I used the following research methods and tools: qualitative interviews, systematic survey of the server's contents, automatic metadata extraction from files to prepare a metadata timeline, narrative of the archive's formation process, and consultation of the institution's accounting books and records.

Qualitative Interviews

Working and technical context data came from the interviews to 16 former staff members and 4 IT consultants who worked for the institution in various periods. The number of staff members interviewed represents 70% of those whose records were stored in the shared directory on the networked server. I created two interview protocols (See Appendix I: Interview Protocols), one focused on work practices, including the types of IT tools used to create and store records; the other one on technical issues particular to

the IT development in the foundation. Because some interviewees were relevant in both contexts, I used both protocols to question them.

The interviews were unstructured. Depending on what the participants remembered, what their specific function was, and at what point they worked in the organization, I added questions and modified others on the fly. For example, some of the employees who worked at the beginning of the foundation remembered more about the paper record-keeping practices, others did not use the network, and only a few were involved in the IT decision-making process. During the interviews I learned about the different records-creation and records-keeping practices, how subsequent information technologies were used, and who had been involved in selecting and implementing them. The interviews provided insight about work-practices and the relationships between organizational functions, staff members, use of the databases, and types of records produced. I found that most interviewees did not remember specifics about the information technologies used over the last 20 years, such as names, dates, and versions of hardware and software. Instead, the interviews with IT consultants and former systems administrators contained details about the technologies used in the development of the foundation's databases and for everyday operations, as well as about the way in which they were implemented and maintained.

I tape-recorded the interviews and gave the interviewees' a false name to disguise their identity. For the purposes of analysis, data points were transcribed and organized in two data sheets: a staff timeline includes name and dates of employment for each staff member whose records were stored on the shared directory and the dates covered by their records; an interview sheet contains staff members' job title, changes in roles, functions, other staff members they worked with, types of records that they created, records-creation

and record-keeping practices, and use of IT. For each interview I noted the interviewee's "name," the cassette number, and the interview starting time within the tape. While the data sheets were extremely useful for many purposes, I had to listen to the interviews at least three times to capture details that I wanted to include in the dissertation and wished I had transcribed them. I used the interview data to construct the narrative of the archive's formation process, to construct parts of the metadata timeline, and to validate the results of the appraisal method.

Technical survey of the networked server

To learn more about the technological context I conducted a systematic survey of the networked server. Using an inventory form, I recorded the names of the upper level directories, their provenance (functional area or creator), encompassing file dates, whether they contained records or applications, and if it was possible to render the files or to run the applications. Within the directories, I studied the files as material culture. That is, I considered the files belonging to records, systems, and applications as artifacts by looking at their format, their properties, dates of creation and modification, creator, and relationship with the other files. In many cases, the contents of the directories were easily identified; in others, I had to search for information on file extensions in different online sources, process them for automatic metadata extraction, or ask for help from the former systems administrators.

Within each upper level directory there were many other nested sub-directories which were analyzed to determine whether they where holding applications or records, to determine the function to which these belonged, the types of file formats included and their date ranges. And in the case of sub-directories containing records from the staff members they were analyzed to determine individual record-keeping practices. The data

obtained was used in the narrative of the archive's formation process and to create the metadata timeline. Both sources were used for the digital appraisal method and the preservation strategy.

Automatic metadata extraction

Another way to obtain information about the technical context was through automatic metadata extraction from samples of text and spreadsheet files belonging to the staff members who worked in the organization from 1991 to 2005. Using DROID (Digital Record Object Identification),[46] a JAVA based file format identification program created by the National Archives in the UK, I identified the names and versions of text editing software used in the institution from 1997 to 2005. The limitation with this tool is that for text files created with non-Windows compatible software, it offers only tentative results. Results from DROID are output in XML and I exported them to an Excel spreadsheet for analysis. Since up until 1997 the institution was moving gradually from DOS to Windows and different people were using different text editors and operating systems, I used the FileMerlin™ [47]conversion tool to confirm the software name and version of files created prior to that year. Furthermore, to learn how many people were using what type of software and when, I sampled files from each staff member represented on the shared directory for a given year. File identification data informs the narration of the archive's formation, the metadata timeline, and the preservation strategy.

Archival sources of IT information

I found information about the IT enterprise in various sources in the institution's archives. Electronic records found in some of the staff members' folders contained inventory lists of the computer models used by each of the members of the institution. These pertain to the years in which the institution hired a part-time systems administrator.

I also found paper versions of these lists in the IT office. The legal accounting book, in which major computing purchases were recorded with a very general level of detail, constituted an accurate source to learn about dates in which equipment upgrades were made.

Printed versions of the annual reports and electronic records from the shared network were used as sources about the institution's background and to frame the natural archive formation process within the foundation's evolution. During the process of archiving the records, I wrote proposals, letters, processing manuals and instructions, memos and reports that I mention in this dissertation. For confidentiality reasons, many of these documents cannot be referenced or included as appendixes, as the real name of the foundation and its internal affairs would be disclosed. I do include the Transfer and Maintenance Protocol that I prepared as part of the preservation strategy after changing the name of the foundation and of the IT consultants to which it was addressed.

Metadata timeline

The metadata timeline is both a research product and a tool. Using data gathered from the interviews, from archival records and the legal accounting book, from online sources, and from the metadata extraction document, I constructed a document that contains hardware and software names and versions used in the institution from 1987 to 2005. It is included in Appendix II. Where I could find the information, I included both the dates of commercial release and of product discontinuation, as well as the dates of implementation and final use at the foundation so as to perform cross-referenced analysis of the broader context in which technology decisions were made. This document was useful for writing the narrative about the archive's formation and was indispensable for making preservation decisions.

PART II: PAST

A Case Study

In the last five years the focus of archival case studies has steered towards electronic record-keeping systems. As the most representative example, the InterPARES 2 case studies explore the meanings of records' authenticity to different communities of practice that use electronic systems to create and maintain electronic records. The goal of these studies is to understand the nature and uses of these systems to define best practices and standards for future electronic record-keeping systems and electronic records.[48]

The study of Aleph's electronic archive differs both in its characteristics and in the purposes of the research. It focuses on an electronic archive created and developed during the late 1980s and 1990s, a twenty-year period during which information technologies were being massively adopted in the workplace. At this time there were no rules regarding what an electronic record-keeping system ought to be, nor any preconceptions about what electronic records meant to their creators or to society. Instead, there was an established tradition of paper record-keeping practices and set values associated with paper records. At Aleph, the coexistence and blending of electronic and paper systems to carry out transactions and to store records generated tensions between the two formats that went unnoticed until a decision about the institution's archive had to be made and the value of the electronic portion came into question. Thus, a salient characteristic of this study is that it looks at electronic records in contrast to the paper records not as antagonistic to them; as belonging in short to the same archive.

Two goals of this study are to establish the evidential value of the electronic records of this archive and to find a strategy to preserve them. Therefore, there is a dual emphasis on understanding their social uses and functions and their technical dependencies. Just as the paper portion of the archive lived in its system of folders and file cabinets, the electronic records were stored in a networked server surrounded by other systems and applications that enabled them technically and complemented them from a work perspective. For instance, printer drivers allowed the records to be copied onto paper, and the database systems generated information that became the content of some of the records, all of which adds to the evidential value of the records as well as to the complexity of preserving them. Ultimately, this case study shows how the characteristics and scope of this archive led me to define the concept of "natural electronic archive" that, in turn, applies as a category of archive to other electronic archives generated under similar conditions.

FORMATION PROCESS AND DIGITAL CULTURAL MATERIAL

Having conceived the networked server as a digital archaeological site, I frame its study in the likeness of a site formation process. Michael B. Schiffer describes archaeological site formation processes as "the pathways of material entities from any past activity to the present."[49] Studies of site formation processes follow material culture as it traverses time, going through functional and physical changes caused by people and by the environment. Their outcomes emerge through scientific and physical examination of artifacts; studies of how artifacts were used and about the environments and relationships in which they were found; and from inferences based on secondary sources. This research approach has been applied by Material Culture scholars to study

archaeological remains, private and museum archaeological collections, and anthropological archival records.[50] Most recently, the study of formation processes has been used to study technological artifacts that; whether buried in land-fills or stashed in somebody's basement, are defining products of human activities since the late nineteenth century.[51] Underlying formation process studies is the possibility of establishing the authenticity of the artifacts from which conclusions and interpretations are drawn. The more that can be learned about the origins and evolution of artifacts and how their use and symbolic meaning has changed over time, the more it is possible to ascertain their authenticity. In archival terms, authenticity is, "the quality of being genuine, not a counterfeit, and free from tampering, and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context." [52] Specifically in the case of a digital archive that lacked explicit rules regarding its use and maintenance and for which documentation was scarce, formation process studies can mitigate or close documentation gaps. And yet, it is legitimate to ask if it is reasonable to consider digital files as material culture.

The "materiality" of digital media is a current discussion in the field of Digital Humanities. For example, storage devices holding electronic texts and gaming files are explored with digital forensic tools. In relation to preservation, the discussion questions whether digital media should be considered "vulnerable," "evanescent," and "ephemeral" —all conceptions that derive from a romantic view of the media and imply the impossibility of preserving—and whether the complex coding and computational processes through which digital objects emerge can be ignored. The opposing point of view sustains the argument that hardware and software materiality can and should be explored, emphasizing more the possibility of preserving digital media.[53]

In this study I treat files as artifacts with the purpose of studying the formation process of an electronic record-keeping system of the past, not an archive.[54] I use the concept of material culture to show that computer files can be examined closely; that time, use, and the environment leave specific marks that can be accurately interpreted. There are several ways of observing older digital material, of which I will mention the few and fairly simple ones that I used in this study. Depending on format specifications, properties embedded in files can be viewed and extracted to identify files' format and version, examine dates of creation and modification, and often even to disclose the name of their creators and subsequent editors.[55] Some old formats can be viewed with file viewers, or can be run through their native programs or emulators to determine their behaviors and what they looked like. Specifically in the case of Aleph's archive, studying the digital elements surrounding the records on the server, including hardware drivers and programs, allowed me to learn about the tools used to manage, display, process, and modify them. In addition to the notion of the archaeology of the server's formation process that I use to approach this study, the use of technology to understand technology has itself come to be known as "digital archaeology."

Furthermore, the social and work contexts surrounding these records provide the clues to the reasons and circumstances through which they originated and why they survived. Among those circumstances, the effects of my research and my work archiving Aleph's records have to be acknowledged as part of the formation process of this archive. My research was situated in the midst of the foundation's closure, intertwined with the archiving process, which was directly influenced by the pressure that I brought to bear on its creators to preserve the archive.

At this juncture, the concept of reflexive archaeology allowed me to work with these different and sometimes conflicting interests, and to go back and forth from the archive's past to the needs and constraints of the present situation. Following the precepts of reflexive archaeology, my study incorporates the perspectives of stakeholders including me. It relies heavily on the interviews that I conducted with the records creators and highlights the testimony of the staff member who initiated the computing enterprise and played a major role in the records-creation and record-keeping practices in the organization. It focuses on how the electronic records were used in the context of work-practices, habits, institutional policies, local legal requirements, and in relation to the established paper system and paper tradition. The narration of the archive's formation process is also a consequence of considering the networked server as an archaeological site, attending to its role as a source of information and as a validator of other sources. Similar to the work of digging a real archaeological site for which horizontal plans and vertical sections are drawn to reveal the stratigraphy of actual superposition from the deposits of sequential events, it is necessary first to locate the server's contents in time, and to explain why they are relevant.

THE SERVER AS A SITE

At the center of the case study is the networked server, or more accurately, the digital material sustained in the server's environment, some of which dates back to the beginning of Aleph's computing enterprise. Grouped in broad categories, the server contents are as follows: staff members' electronic records since 1991; backups of email mailboxes belonging to some staff members since 1998 and the enabling script for those backups; vestiges of the first and second database systems and data files from 1987 and

1991 respectively, the last database version having been in place since 1997; and a variety of applications and drivers dating from 1987 to 2005. These digital objects correspond to two distinct personal computer operating systems, DOS and Windows. Over the years, staff members and systems administrators deposited these digital materials, which in turn were transferred from one server to the next and from the hard-drives of staff members' work-stations to the servers. This research focuses on the study of the database systems and the electronic records deposited in the shared directory on the networked server.

As a stand-alone primary source the server has both limitations and advantages. As for its limitation, again the analogy to an archaeological site comes to mind, since I can only count, observe, and analyze what is there, but I have no way of precisely knowing what was deleted, purposely or accidentally, placed on other storage devices, or lost when old equipment was abandoned. For instance, it is unknown whether some employees removed their records before leaving the institution and whether applications were removed over the years. However, when combined with the interview data, some unknowns are revealed. Two staff members interviewed commented that they consolidated files from their predecessors in their own directories, which explains why one of the folders for an earlier staff member is not present in the shared directory and the other was buried in an unlikely sub-directory.

Among the advantages of the server as source, files can be identified and studied to determine format, dates, and software dependencies. Being able to study electronic records and the tools used to create them in parallel provides context for the electronic records and raises their evidential value. Not only do the records' contents and the record-keeping structure provide validation to an archive, but technology can also

validate the records' form, period and type of use, and in some cases even their provenance. What follows are two examples of how the analysis of these materials allows reconstructing in the present a recordkeeping system of the past and speaks to the role of the server as a research validator.

A Hewlett Packard LC3 server with Pentium III processor running the Windows NT operating system purchased in 2000 was the last file server used at Aleph. Located in one of the directories I found .prg and .dbs files belonging to Aleph's first custom-made databases.[56] With a text editor I was able to read the database functions' programs, and I opened .dbs files with Excel and with one of the many dBase file viewers downloaded from the Internet. Fortunately, a few programs had date and version annotations which, along with their last modification dates embedded in the file properties, allowed me to establish that the development of the first database system dated back to 1987, thus resolving a difference of two years between the dates provided by two interviewees.

A similar situation enabled me to place the implementation of the second database system in the year 1991. In the winter of 2005 while I was conducting one of the interviews, the interviewee opened the second version of the database system from a desktop icon located on his computer running Windows 98. While he was able to execute some navigation commands successfully, and even though this system was DOS compatible, he could not pull up any data. The path in the desktop icon led to a folder on the networked server. The analysis of the files included in this folder helped me figure out inaccuracies in the dates provided by different interviewees. The reason why vestiges of these two early systems survived from 1987 until 2006 can only be inferred. Despite changes in the design and platform of the database systems, the data was always transferred seamlessly from one system to the next.[57] Possibly, program files and

executables were transferred to the new system during these migrations for testing or backup purposes, and were never removed by the succeeding system administrator, who preferred to retain old files rather than to delete them.[58]

None of the interviewees remembered precisely the moment in which the shared directory was implemented on the networked server as the storage and exchange location for their files. One interviewee indicated that it could have been circa 1992, and another one remembered that at some point after he started working that year, somebody had said that they should store their records in their assigned folders in the server. This approximation concurs with the period during which the second database system was put into effect and with a significant purchase of PC equipment.[59] Most significantly, the earliest (and very few) files in the shared directory date from 1991, which brings closer the possibility of establishing its implementation circa 1992.

Files present in the staff's shared directory show that from 1991 to 1993 only three staff members had their records there. From 1993 to 1996 the number of participants in the shared directory increased to six and almost doubled from eight to fifteen from 1997 to 1998. Over the years, the maximum number of staff members with folders on the shared drive was seventeen in 2004. The gradual appearance of participants in the network could be due to some of the following reasons: that in the early years not all the staff members used computers, that at the beginning not all the computers were connected to the network, that there was not enough space to store everybody's files and people used their hard-drives or floppy disks, and that some employees deleted their files upon leaving the foundation.[60] The way in which the increase occurred, reaching a peak and stabilizing in 1998, is consistent with the evolution of computing at Aleph. Between 1997 and 1998 the foundation moved to a Windows environment, the third iteration of

the database system entered the production phase, more work-stations were upgraded to Pentium, a Sun Microsystems server for Internet access was purchased, and the local network for PCs was improved.

The examples above introduce the role of the server as a site and highlight the main elements used to study this case. In the next section, the narrative of the archive's formation process combines the server data, secondary sources the foundation's records, and interview data to reveal the evolution and uses of the electronic and paper systems. Throughout the foundation's tenure the paper and the electronic systems converged and diverged, and in these comings and goings, the meanings and uses of electronic and paper records changed as well.

The Paper Path

PAPER FILING AND WORKFLOW

Renee worked for Aleph Foundation from 1985 to 1993, first as the institution's general secretary and during the last two years of her tenure supervising grant projects' progress and grant payments. As a secretary she did a little bit of everything: distributed correspondence, typed letters of rejection and acceptance to applicants, supervised projects' progress, and wrote reports and board meeting minutes. She also had a small group of assistants. Due to her prior experience as an executive secretary in the private sector she was charged with "inventing the archive."[61]

While not labeled as such, during the first seven years of operations the foundation had a centralized record-keeping system. The system devised by Renee was adequate for the new organization, which dealt with a small number of requests on a first come, first served basis. It consisted of a records work-flow, a filing system, and a

register. In general, applications, petitions, correspondence, reports, plans, and memos circulated among all senior staff members. Each record was accompanied by a sign-off sheet on which staff members could write opinions and recommendations. If a record was circulated only among certain staff members, Renee sealed it with the name of the staff member to whom it had to be directed. The filing system comprised three sub-systems, each with its correspondent alphanumeric coding. A general subject file contained records related to all the issues handled by the organization, a chronological file included copies of incoming and outgoing correspondence which were also filed in the subject file, and a project's file contained all the records corresponding to each approved grant.

To track projects Renee kept a notebook-scheduler in which the active projects' names and numbers were listed, including the dates on which progress reports from beneficiaries were due. Every 15 days one of her assistants would go through the notebook to verify that reports had been received and request those that were overdue from the relevant beneficiary. As the number of approved projects increased, going through the scheduler to track progress became "a pain."[62] All the records were controlled through a registry that indicated their circulation and location at any time and were filed every day at 4:30 pm before operations ended at 5pm. Renee indicated which ones had to be photocopied and where to file them. The efficiency of the system was described by Pedro, the foundation's long term executive director.

> *At the time, we ran the foundation with electric typewriters (with short magnetic memories), a telex machine and the first fax machines. We kept our files the usual way: papers in folders and filing cabinets. Efficient secretaries took care of that. I did not interfere (in fact, had little or no interest in the matter and trusted that somebody took good care of papers, which were always produced in good order when required).* [63]

By the time Renee retired in 1994, the foundation had achieved its final organizational structure, which included the areas of Arts and Cultural Heritage, Social

Welfare, Education and Sciences, Administration and General Services, Executive Direction, and Presidency. It had also consolidated its lines of grants and was receiving an increasing number of applications. Computer technologies had been in place since 1987 to track grants and finances, and almost everybody was using word processors and stored their records in a shared directory on the networked server. Some of the former secretaries became project coordinators, and new staff members were hired to work in the different areas. As a consequence of the expansion and of the diversification of functions, some employees started to keep area-specific records in their office spaces. Gradually, the general subject and the chronological files created by Renee faded out as various record-keeping systems emerged and changed with the hiring of the next new employee.[64]

Natalia began to work in mid 1997 as project assistant in the Arts and Cultural Heritage area. She found that the centralized subject and correspondence files that were maintained by her predecessor did not have "direct relationship with anything,"[65] and decided to create her own subject categories reflecting the programs that were current at the moment. By contrast, Eliana, the long term president's secretary, continued for her area's records with the traditional filing system implemented by Renee.[66] Regarding what was kept and discarded, the exception of administrative and board records that were filed systematically and following records retention requirements for tax and accounting records,[67] everybody else made their own decisions. In general, staff members gauged items according to the priorities and needs of the moment and tended not to discard much, as they frequently found that shortly after they threw something away they needed it.[68] The only file system that remained as it was originally designed until the end of the foundation's existence was the projects file.

The projects paper file is not complete. The basement where it was stored suffered two floods, and damaged records were discarded on both occasions. To make space on the shelves for incoming folders, older ones were disposed of on an irregular basis. Weeding was done on folders from the area of Science and Education, whose calls for grants reached up to 6,100 beneficiaries. Juan, who acted as cultural manager for ten years, did not approve of weeding and ordered a considerable amount of files to be moved into his area's work-space to avoid it. In all cases, the "mother" folder containing the production documentation for the call for grants and the lists of beneficiaries was retained, but all the "sons" folders with the progress of each individual beneficiary were discarded. There is no record of what was lost in the floods, and only random notes with lists of files remain from the weeding sprees.

WRITING AND TYPING

Like the early record-keeping system, the record-making practices in the early years also reflected the executive secretary style. Documents were hand-written by senior staff and given to the secretaries, who typed them using electric typewriters with a small amount of memory, photocopied them, filed a copy in the general subject file, and circulated the original.[69] The first cultural manager, who worked from 1987 until 1992, remembered the drill,

> *The first time I had to write reports was at the foundation. I had a secretary, that poor thing,…she had to understand my handwriting, type the report up, and then I had to review it and re-do it. It was terrible.* [70]

Initially only the executive director, the area managers, and some administrative employees used word processors, and the secretaries continued using memory typewriters.[71] According to the legal accounting book, the foundation purchased sixteen

personal computers between August of 1991 and June of 1993, and acquired seven more in 1994 and 1995.[72] By then, the foundation had completed ten years of activity; it had achieved its final functional structure and had identified its target interest areas. The number of full time staff members was seventeen. [73]

About the types of text processing software Pedro said,

> *We started using three different word-processing programs: Word, Word Perfect, and WordStar, and for some time each was free to chose which, until we collectively realized that we needed to agree on one and, again a bit by chance and luckily, picked Word.*[74]

Those word processors must have been used prior to 1991. According to the study of file formats found on the shared directory, in 1991 the foundation was using Microsoft Word 5.0 for DOS. During the next few years until 1993 they continued using that software as well as its 5.5 upgrade. From 1994 through 1996 both Word 5.5 for DOS and Word for Windows 6 coexisted. By 1997 all the work-stations were using versions 6 and 7 of Microsoft Word for Windows.[75] Licenses for operating systems and software were purchased individually for each computer station.

Document types developed in tandem with the organization. Federico, the first social manager, suggested preparing an action plan, a document in which the different managers laid out their strategies and their budget needs for the next fiscal year. The foundation's president, Jose, initiated a format of board meeting agenda drawn from his experience in the corporate world. The foundation policies and the grants' general conditions, which constituted the road map of the foundation's decisions and the beneficiaries' responsibilities, were written by Pedro.

From the early years in which secretaries typed their supervisors' manuscripts almost until the institution closed, all the official documentation was edited by the executive director. About his roles as records creator and editor, Pedro said,

> *As Mr Jourdan with prose, I was not aware of having a "records creation routine." I simply wrote stuff. Much of it was the foundation's formal communications materials, such as published policies, standard letters to applicants accepting or rejecting their proposals, conditions of entry for open calls for applications, ad hoc letters in case of non-routine proposals, etc. I was particularly keen on making sure that all communication with the outside world was very carefully worded: regardless of who actually signed it, all documents issued by the foundation required my approval (as an editor, one could say). Many of these documents came to my desk as first versions written by others for me ("that pestilent fellow, the critical reader," as Henry Fowler said) to edit. All board-meeting agendas and minutes were also written by me. And, of course, I wrote or edited any letter of agreement or understanding with outside parties.* [76]

Over time, these document types became electronic templates that were filled in with new names and dates and topics over and over by different members of the organization. Beyond the frustration that the editing may have caused to the employees, during the interviews almost everybody mentioned that they had learned from Pedro's corrections and writing style and some mentioned that they were doing the same in their current jobs. The imprint must indeed have been strong: during our meeting in 2005 Valentín, who left the foundation in 1992, recited the beginning of the model letter of acceptance and/or rejection by heart.[77]

The Computing Path

INITIATIVES

The initiative to computerize operations was driven by Pedro, the executive director.[78] The transcription of part of his interview describes the beginnings of the enterprise.

> *I had some previous experience in computing, first as an academic in the 60s and then as a consultant in the 70s. The original general manager bought one of the first models of IBM PC on the market, probably in 1987 or 88, to keep his*

financial information and help him make decisions. When he left I inherited the blessed machine but had no clue on what to do with it. Around 1989, if my memory doesn't fail me, we decided to explore IT tools to help us with our day-to-day business.[79] *That seemed to be what had to be done at the time in any self-respecting institution, even though we did not know why. I liked the challenge and took the matter directly in my hands, starting by lengthy interviews with half a dozen hopeful consultants with whom to share my carefully (and unsuccessfully) hidden confusion. Most recommended that we buy a large mainframe IBM contraption and set up a network of "dumb" terminals. An elderly physicist, who came from academia, recommended a PC network. With the benefit of hindsight it is easy to see he was right, but most thought otherwise at the time. I must resist the temptation of saying now that I ruled in favour of the physicist because it was clear in my mind that computing was moving in that direction. This, of course, is nonsense: probably the real reason was that what he said made sense, that it was fun to try something new, that I liked the bloke immensely, that he did not seem mainly motivated by making consulting fees and that his solution caught my fancy, perhaps because it would allow me more room to poke my fingers into the future electronic pie.*

Once the decision to go for the PC network was made, we had to find out what the thing would be capable of doing: we were about to buy a tool the purpose of which we were rather in the dark of, beyond the foggy idea that it would help us better manage our affairs. Our consultant did an excellent job in discovering what we really needed, determining what we didn't require and knocking some sense into our electronically ignorant heads. We should be thankful. We agreed on the fact that we should have two independent but connected software programmes: one to run financial and administrative matters and one to manage grant-making and following-up. The consultant designed and wrote both programs (no adequate canned stuff at the time) using one of the computer languages of the moment (forgot which). He also lent a hand in teaching all and sundry how to operate the computers (we had clones built to measure, except, probably, the server, which also proved to be a reasonable solution, at least in terms of value for money).[80]

Recommending a network of personal computers was certainly bold in the Argentine scene of 1987. Only four years earlier, in 1983, the PC had not arrived yet in the country and IBM mainframe equipment dominated the market. In the following years the first microcomputers arrived in Argentina, but their dissemination was dependent on foreign developments and products, and there was little experience with the technology. Also, imports suffered due to the inflationary crisis following the adoption of the new

currency “austral,” intended to stop inflation, yet meeting with no success. Consequently, equipment availability was erratic and dominated by monopolies and by the black market.[81] But it was not by chance that this consultant offered an innovative alternative to the foundation’s computing concerns. Early in his career he had worked under the direction of two Nobel Prize winners in Nuclear Physics at the University of California at Berkeley and at the University of Paris; in Argentina he was the person people consulted on use of computers in research and teaching.

Among staff members the adoption of new technologies in the late 1980s elicited responses that went from detachment and resistance to excitement and opportunity, as for many it would be their first opportunity to use computers[82] Valentín remembered a funny story that shows the magnitude of the novelty. During a visit to the foundation an academic, member of the board of directors, asked the executive director, “¿Booteaste? (Have you booted?), meaning “Have you learned how to walk?”[83] In 1985 Mark D. Larsen stated that, “no more than a decade separated U.S. from Latin America when it comes to computers.”[84] Aleph is an exemplary case of this assertion. IBM PCs with 286 processors and the Novell NetWare 286 technology had been released in the US five years and one year respectively before the foundation used them in Argentina.

For the following 18 years, decisions about IT developments at Aleph remained under the responsibility of Pedro. It was his job to weigh inputs and opinions, consider financial implications, and secure the appropriation of funds. Gradual improvements were routine matters; major overhauls of either software or hardware went through consultation and planning processes. Changes were implemented by the financial manager with the help of IT consultants and the part-time systems administrator.[85]

SYSTEMS

Databases

Around the time that he was consulted by the foundation, the developer of the first database system traveled to an international conference on Novell NetWare, an operating system created in the US in the early 1980s that allowed multiple tasks to run on networked PCs.[86] When he returned, he volunteered a team of colleagues—all University of Buenos Aires professors—to develop and start running two separate databases, one to manage grant projects and the other to manage finances written in dBase III for DOS and running on Novell NetWare 286.

The team's first challenge was wiring up the building while honoring the request of the architect that the harmony of the building's historic features should not be disturbed. But when the wiring was installed through the attics they found that the heat from the pipes damaged the cables, and the whole circuit had to be reviewed. Once the initial problems were solved, PCs were installed in some offices and the team started interviewing the staff members who would use the systems to learn what data had to be recorded.

The design and implementation process brought mismatched expectations as well as opportunities to learn. In the beginning, staff members asked the consultants to replicate the existing paper workflow to register, tally, and track applications and projects. Gradually they realized the potential of the system and asked for more complex tasks such as adding comments about the projects' progress and different types of statistics. Program files from 1989 and 1990 show that the systems had a considerable number of routines and could produce reports in English and Spanish.[87] Still, the results were not completely satisfactory. Novell NetWare 286 was clumsy and slow, requiring

various adjustments and reboots to get it to work with the DOS based database.[88] I infer that it must have been hard for the consultants to work with a system with such inherent problems, for the users to understand the technical limitations, and for both to deal with their own learning processes. Up until the end of their contract, the consultants remained in charge of managing the systems, making changes requested by the foundation, and generating graphics and statistics needed for annual reports and board meetings.

In 1991 a second IT consulting firm was hired to revamp the grants database and to maintain the computer stations. The custom financial database was discontinued and a standardized product, purchased from a multinational accounting consulting firm, was used instead. Mario, the IT consultant, commented that compared to the banking and commercial systems commissions that his firm was getting at the time, developing a grant tracking system for a philanthropic organization was a novelty. Hardware and software wise, the new system was an upgrade of what was already in place. More computers were purchased and Novell NetWare was upgraded to 386. This was a major improvement, as the new version allowed more memory to be allocated for the database functionality. The new program was written using Clipper 1987, a popular compiler for dBase III. Given the fairly small amount of data that was managed and the resources that could be invested, these technologies were adequate. In those years DOS and Windows systems coexisted and while the market was moving towards the latter, developers were still releasing DOS compliant products, and dBase experts were easy to find.[89] However, the DOS environment was limited in terms of the amount of data that could be entered in the fields, and it did not allow text processing functions. Due to the disconnect between the grant tracking and the financial systems, the same data about each project had to be entered twice, requiring two full time staff members in the administrative area to do the

work. This disconnection also implied that the administrative staff did not know about the evaluation and progress of the projects. In late 1996 a new IT consultant was called in to redesign the system and to combine functions of the financial and grant tracking databases.

The system's third iteration was written in Clarion 2.2, a Windows compatible fourth generation programming language for databases. The development lasted eight months, after which the data from the second system was migrated seamlessly during a weekend. The advantages were significant. The financial system was customized with input from the financial manager and following standards set by the umbrella organization that consolidated the information of the three sister foundations. In the new system, payments to beneficiaries could be authorized upon the receipt of progress reports from the grant tracking side of the system and were automatically deducted from their accounts at the financial end. Also, the interface was more user-friendly, providing different ways of entry. The amount of text that could be entered allowed a better follow up of the beneficiaries and the letters of grant offer—whose wording and style had been pre-established by Pedro—were generated as Word documents directly from the system.[90] In 1998 the foundation purchased a new NT4 Windows server and stopped using Novell NetWare.

Access to the grant tracking system was available to all staff members through individual passwords. The system became a major tool for project coordinators and assistants, who used it for general references, to register applicants, generate project numbers and letters of acceptance, schedule and authorize payments, track projects' progress, and verify grant completion. On the other hand, the only employee with access to the financial end of the system was the financial manager, who was also the only one

who could rectify errors in payments. As a consequence of these changes, the two financial data entry clerks were not needed any longer and the amount of work increased on the projects side of things. Around the dates of the major deadlines for grants, project assistants and receptionists would spend up to 15 days entering data as they received the applications by mail or in person, and many times part-timers were hired to assist.

Systems administration

As day to day computing demands increased, in 1998 a part-time systems administrator was hired to deal with network administration and security; software and hardware purchasing, installation and maintenance; communicating with the Internet provider; and overall cooperation with IT initiatives. Carlos had not completed his degree in systems administration yet, and he learned in the process of doing. He was not given major indications about what and how to do things, and while he formally reported to the financial manager, all the decisions were made in consultation with Pedro. IT procedures grew gradually as time unfolded, on the fly, by trial and error, and based on recommendations from the IT consultants and the Internet provider.

Among his duties, Carlos kept an updated registry of current software licenses and produced computer equipment inventories to make sure that no illegal software was ever downloaded onto the machines. On his own initiative, he decided not to discard any unused hardware, software, or storage device that he found when he started his job and during the almost seven years that he worked for Aleph. He followed the same practice with older files and applications stored on the networked server. Moreover, when a staff member left the foundation, if he or she had left files in their work-station's hard-drives, Carlos copied them onto CDs and also transferred them to the networked server.

One of his contributions was to improve the networked server backup procedure which before he started to work at Aleph was only performed for the grant tracking database onto a zip drive. A tape-based backup set was purchased and implemented in the typical disaster recovery business practice; a set of cumulative tapes one for each day of the week would be re-recorded over the next week so that only the latest data could be recovered. The hard-drives were never backed up. Over time, more servers were incorporated into the network, including a firewall server and a corporate anti-virus server. Close to the end of the institution's life a spare server from a closed project was dedicated as a mirror to the file server, albeit with the downside that both were located in the same room.

The third IT consultant continued providing database systems maintenance and upgrades and website development and management until the institution closed in 2006, and then in July of 2007 assisted me in the archiving process. The association of the consultants with the foundation was based on mutual trust; there were never contracts, and work orders were based on verbal agreements. System copyrights or the confidentiality of the data within them were never discussed. Major systems developments were fee-based, and day to day maintenance of the systems was performed under a monthly subscription. As for staff training, the foundation had the approach that it was the business of all employees to learn what they needed in a reasonable time and that they would give specific help to anybody who asked for it. The IT consultant did not produce (and was not requested to do so) documentation about the system or a user manual. When the third database started running he provided basic instructions to the employees, who in turn helped new employees as they came in. Some staff members wrote up brief instructions for themselves, but these were not shared. It was characteristic

of the organization's culture that everybody had his own way of doing things, and the system's capabilities were known by some better than by others. The grant tracking system became the essential tool of project coordinators and assistants. Some managers used it heavily, and others did not use it or did not know how to operate it. When the latter had doubts or wanted to learn about a project they asked their project coordinators to look into the subject and discussed it with them.

Internet: Website and Email

The foundation's first website dates at least back to 1996, although it could have been launched earlier.[91] The idea of creating a website came from Juan, the cultural manager who was interested in the use of the World Wide Web as an artistic medium and brought in an artist as the first designer. This was his first and final involvement with the project, which was subsequently managed by Pedro and carried out by his assistant first and later by Diana, the education manager. The technical aspects of the website and its management were always subcontracted; the content was supervised and edited by Pedro and compiled by the staff member who delivered it to the webmaster as a Word document.

Through the years, the website evolved from a static HTML page to a Java scripted one in 2002, which was when the IT consultants were commissioned along with a graphic designer to create a new website. When Diana, the Education Manager, was assigned as a liaison between the foundation and the webmaster, to avoid misplacement of content she sent updates along with a document indicating their location on the site. Updating involved uploading and replacing old files with new ones, and the contents were never stored consistently. There is no complete record of the foundation's website; only .doc files scattered in the directories of the staff members involved in the sites

workflow. Samples of the website since 1996 can be recovered from the Internet Archive.[92]

Internet services including email and website hosting were provided by a nonprofit academic telecommunications network agency with which the foundation had a 1 megabyte (Mb) broadband exclusive point to point connection.[93] Email mailboxes remained on the workstations' hard-disks and were managed by the staff members themselves rather than centrally. In 1999, the systems administrator created a script on the networked server to backup some of the email accounts that crashed very frequently. Not all the employees were aware that this security measure existed and that their backup email mailboxes on the networked server were potentially accessible to everybody. Increasingly, email became the preferred mode of communication, so much so that in 2001 the foundation decided to stop sending the forms for the call for grants through regular mail, and the receptionists became responsible for distributing them by email.

Finale

Anticipating the end of Aleph's activities somewhere between 2004 and 2005, in 2002 the foundation decided to reduce computing equipment updates. The decision was also hastened by the Argentine economic crisis of the year 2001 that tripled the price of the dollar and the costs of computing equipment. Most hardware and software in the institution did not evolve beyond Pentium II and III with Windows 98 as operating system (OS) and Microsoft Office 97 as the standard desktop program suite.[94] That same year the IT consultants were commissioned to add a web-based form to the grant tracking system to allow online registration for the calls for grants, and to automate the administrative process that followed the selection of beneficiaries. The web registration system was test implemented in 2003 and used throughout 2004 to manage the last calls

in the area of Science and Education, believing it was a community familiar with the use of technology and with online applications. The coordinators in the areas of Arts and Social Welfare, however, decided that it was going to be cumbersome for their audiences to adjust to a new system and that it was not worth attempting the trial for the last call of grants.

The addition allowed Science and Education applicants to register online through a password protected interface so that their data could be entered automatically in the grant tracking database. Internally, the project assistants, the area manager, the administrative manager, and the executive director had different levels of authorization in a process that went from application review, to approval of funds, and finally to automatic emission of an email offering the grant. To retrieve the letter of approval—which was generated with a unique number—the recipient had to go back to the password-protected online interface and accept the grant conditions.

And yet the process did not go totally paperless. Applicants still had to send background records through regular mail and paper forms were available for those who did not have computers or email accounts. The option to allow applicants to submit résumés and letters electronically was discussed, but it was agreed that downloading and printing materials that had to be passed on to the evaluators as paper was going to overload the staff. Once the system registered the applicant's acceptance, a copy of the electronic letter of approval was printed and included in the project file. Considering Argentina's bureaucratic culture and the importance of signatures and seals, Diana expected many requests for signed certifications that recipients had been awarded a grant. Interestingly, only 10% of the awardees asked for such letters.

At a time during which the foundation was downsizing, the new system reduced data entry as well as printing and mailing letters of approval. Whether the reason was that these were the last grants given by Aleph, or because of the ease of the technology, the number of applications rose significantly. When I asked Pedro why the foundation decided to implement this system so close to the end of its existence, he explained to me that—besides the practical aspects of reducing personnel—it was a way of denying the proximity of death.[95]

THE SHARED DIRECTORY

> *On the server's hard-drive there was a folder for every record-creator, identified by their name/initials: within the folder, each did as well they pleased. Again, in a small community, such freedom worked well.* [96]

Implementing a shared directory on the networked server to store electronic records was Pedro's idea so that everybody could see and access each others' files, and paper did not need to be circulated as much. In the shared directory, each staff member had a directory named with his or her initials to store their files. Four general sub-directories were also created in which staff members could place final versions of records of general interest such as policies, forms for calls for grants, and annual reports. Lacking explicit records management rules, the purposes and uses of the shared directory were known and interpreted differently by each staff member.

Anahí was a project assistant in the area of Science and Education. Once the final versions of official documents were released and distributed she added them to the sub-directory "Alephdocs" from which she could pull copies that she used as templates to create similar record types. Her supervisor told her about this collection, but she did not know who started it or if anybody else contributed to it. Actually, it was almost at the end

of her seven-year tenure when Celia, the social welfare manager, learned the function of these folders, which she saw every day on her computer screen and never opened, as she followed the path to her own directory. She then thought how she could have used them to create a chronological collection of the different grant application forms issued in her area.

Network uses and perceptions

The idea of accessing everybody else's files was appealing, but the action itself was dominated by prudence and had practical limitations. The following quotations, each one transcribed from a different interview, reflect the fact that opening somebody else's folder was perceived as invading his or her privacy. [97]

> *I don't get involved in what other people do, and I die if somebody changes something that I did.*
>
> *In general you ask permission before going into other person's records.*
>
> *I would never get into somebody else's folder for the sake of looking at his/her records, only to look for something that I need.*
>
> *I interpreted that "Alephdocs" was somebody's folder and I never intrude into my boss's folder, unless he tells me to do so to find something.*

A common practice was to ask the folder's owner before going in looking for a file. On the other hand, finding something in someone else's folder without his or her help—and sometimes even with their help—took a long time or was often impossible due to the lack of uniform record-keeping and file naming convention practices. As recalled by Pedro, "Folder system: yes. Found other people's records as best one could. If all else failed, asked the culprit."[98] Exceptions happened between employees who worked closely together and had an informal but explicit agreement that they could open each other's folders and understood the logic of each other's organization and naming scheme. Still,

even when it involved asking for permission, the shared directory avoided the inconvenience of having to print records or going to the other staff member's workstation to retrieve information.

Network unconsciousness

By network unconsciousness I mean the low awareness that staff members had about the use of the shared directory, its privacy, and its technical boundaries. All the employees had the ability to store records on both their station's hard-drive and in the shared directory. The former was chosen by some employees to store personal records, and the latter was the likely choice for work-related records; a practice indirectly (not explicitly) enforced by the administration. For example, only the shared drive was consistently backed up, although most staff members did not know about this policy.

That some employees kept their personal records on the shared directory suggests network unconsciousness: an assumption that nobody was going to look at them. This connects well with the prudence that governed the use of the shared directory. In general people did not look into other people's records without asking for permission, and this notion overcame the reality that all the records in the directory were exposed for everyone to see. Celia's impression was, "I don't think that anybody intended to look for something in my folder without telling me. If somebody needed something they asked me, and I said, do I show you the path or I send it to you by email?"[99]

That staff members used the shared directory for almost fifteen years without considering that it could have been improved or used differently is another example of network unconsciousness. It was not until our interviews, when I asked staff members how they found each other's records, that some mentioned that if the institution had continued they would have had to find other ways to manage their files. About that

Diana's comment was, "I regret not having organized my electronic records; it would have helped me find a lot of things."[100] Pedro's answer to my question about his electronic records' weeding criteria also reveals the unconscious aspect of electronic record-keeping, "No conscious criterion on keeping electronic records. Instinctively, keep everything. Bear in mind that the amount of information handled was relatively small." [101]

At the highest level of network unconsciousness, the comments of two staff members interviewed indicated that they did not remember or know about the boundaries between the shared drive and the hard-drive nor had ever considered the option of using one over the other. They just used what had been pointed out to them and did not think further. This lack of understanding of the technology was familiar to me: during the time that I used the network at Aleph, I did not give a minute's thought to these issues.

Confidentiality [102]

Above and beyond the staff's perceptions of privacy, the administration did consider issues of security and confidentiality, both with reference to the beneficiaries' information and internally with reference to their own. Even though the foundation's records and data were considered closed to outsiders, most were accessible to all working in the foundation, particularly those records containing information on grants requested, awarded and denied. There was never an attempt to restrict access to electronic information because over the years few leaks happened. There was the impression that they would not have been prevented by limiting access to the electronic files, since evidence usually pointed to indiscreet verbal comments by someone who would have had the information anyway.

That being said, some very specific records were highly confidential and handled by only very few persons, on a need-to-know basis; staff appraisals, salaries, and benefits; some financial documents; or certain correspondence addressed to or sent by senior members of staff or by board members are some examples. These documents were not kept on the shared directory but on the concerned parties' computers' hard disks, which were inaccessible to all not having network administrator privileges (privileges limited to the system administrator and, as a backup, the financial manager and the executive director, who rarely made use of them), and in the case of salaries on a computer not linked to the intranet at all. The foundation became more security conscious with the connection to broadband Internet and installed a firewall and other appropriate measures to prevent outside access.

OTHER DIGITAL OBJECTS

Among other digital objects that I identified on the networked server are drivers for the printers that were connected to the network, two different DOS-based file management systems, database systems created to manage the foundation's library and videos, databases for projects conducted in collaboration with other institutions, images and text records belonging to projects conducted by part-time project coordinators, old programming files, and old files from Lotus 1-2-3 used from 1989 until 1997 when the foundation moved to Windows.

As much as they reflect record-making and record keeping practices, these objects also provide a view of how technologies were embraced in the institution and how they affected its employees. Elena was initially hired as a secretary in early 1988, although what made her appealing as an employee—and positioned her as first candidate in the job

application process—was that she was studying for the career of scientific computer specialist.[103] Using the existing dBase files, she produced different programs to calculate statistics requested by different staff members. About her programming work she said,

> *"At that time I worked with the then financial manger who knew how to obtain the best of me. I programmed a lot for Aleph, I generated the annual reports which were less narrated than the current ones, they had more numbers; for example in the year we will pay so much, we paid this much, we promised to spend this other sum."* [104]

Program files with last modified dates from 1991 to 1993 found in a directory with her initials on the networked server show code to sort grant applicants by surname, project number, geographic location, age limit, and target area to which they had applied. There were also scripts to find number of rejections per area, people who had been awarded more than one grant, and to search and sort through bibliographic data files. In 1994 she was promoted to projects coordinator and discontinued her programming activities. When I told Elena that I had found her scripts on the networked server under a folder with her name, she was surprised. She logged on to the server every day, but because she went directly to her folder in the shared directory, she never came across this folder which was hidden within nested sub-directories with cryptic naming conventions.

Making and Keeping Electronic Records

This section is an ethnography of records creation and use, based on the interviews and show-and-tell sessions with staff members whose stories where checked against the contents and structure of the networked server. Its purpose was finding out not only what people said they thought about the dichotomy between paper versus electronic

records, but what they actually did and how they depended on both to complete their work.

RECORDS AND DATA

What type of electronic records did you create? Was there an electronic record-keeping policy in place? I asked these two questions of each of the staff members that I interviewed. With one exception, everyone answered the first question with a list of record types such as letters, reports, appropriation requests, email, memos, etc., and the common answer to the second question was a confirmation that record-keeping policies did not exist and that everybody invented their own and kept or deleted records according to their judgment.[105] Pedro's responses were different from the rest in that he also considered the database systems as a type of electronic record-keeping system and the data derived from them as records. His answers are transcribed below:

> *Electronic records were of two kinds: (i) electronic versions of written documents of the usual type (letters, memoranda, reports or whatever) and (ii) new records, created ad hoc, that is, to keep information in a different way, both project-related and financial information. The former had authors who signed their names to them and usually filed them both on the central server's HD and on their office computer's HD. Some (such as myself) kept an additional copy in a laptop or a home computer. And for quite some time paper versions were filed as usual, in addition to the electronic file. These records were usually created by standard application software (word processing, spreadsheets, etc). The second type of records (new records, that would not have existed had no electronic system been implemented) were institutional and anonymous (cards or sheets with details describing every project, from initial request to completion, personnel files, accounts receivable and payable, etc). Most of these were only kept on the server's HD and in various cases a paper version made little sense or was impossible, though conventional files were still required to keep documents providing detailed information we had decided not to digitize. Some, being sensitively confidential (salaries, for instance), were either kept in the one computer not connected to the network (in the administration) or in the directors or managers' home computers. These new records were named, classified and identified in whatever ways (name of beneficiary, date, type of project, etc.) the software determined; traditional records received any name their creators fancied (remember that in times of yore file names could only have 8 or 9*

> *characters). On the server's HD there was a folder for every record-creator, identified by their name/initials: within the folder, each did as well they pleased. Again, in a small community, such freedom worked well.*
>
> *No policies: just crossing bridges as best we could when we came to them (and falling into the river a few times). One could say that as time progressed and the original hardware and software were replaced upon becoming obsolete (our tailor-made software was twice re-written, to say nothing of the successive versions of the DOS, Windows and standard applications), policies on record-keeping transpired de facto.* [106]

The conception of data as a record concurrent with the one expressed in this testimony is characteristic of scientists whose work revolves around gathering and analyzing data sets.[107] It should be pointed out that Pedro had experience with "Clementina" the first computer brought to Argentina by the University of Buenos Aires in 1960 and used until 1966.[108]

UBIQUITY

This section illustrates the ubiquity and the diversity of record types present in the shared directory, including style, size, topics, and even languages.

Official records

Given the specified document formats established early in the foundation's history by Pedro, the content of most official records in the organization was co-written by staff from the different areas and finally edited by Pedro. Typical examples were the board meeting agenda and the meeting minutes, the forms for the calls for grants, the action plan and annual budget, and the annual and quarterly reports. Some of these documents were written in English and in Spanish, and some only in English or only in Spanish. Although the workflow could change, the steps for creating these documents were as follows:[109]

- The project assistants/coordinators/managers wrote drafts containing area specific information;
- The project assistant responsible for tracking the progress of grants wrote quarterly reports. Those were reviewed, expanded, or completed by area project assistants, coordinators, or managers.
- Pieces of text from different areas were compiled by the executive director's secretary.
- Some staff members kept their drafts, as they usually contained more information than after they were edited and included in the final documents. Other staff members replaced their versions with the edited ones.
- Final edits, cuts, and additions were done by the executive director.
- Official documents went to recipients on paper or electronically. For example, they were distributed among board members; or printed in the annual report, etc.
- Final and complete versions including the information for all areas were kept electronically by some managers, project coordinators, and assistants, and always by the executive director and the president.
- All final versions were printed and filed, in different project files as well as in different area files.

Considering that board meetings took place once a month, that quarterly reports were prepared every four months, that calls for grants and conditions were issued at least twice a year for the regular series of calls for grants[110] and more frequently for specific programs, and that staff members involved in the creation of these records kept various

electronic versions of the different records, the amount of repetitive information stored in the shared directory was significant. But these were only a portion of records stored on the shared drive.

Work records

Besides those that I call official records, staff members created, received, and maintained various types of work documents. This group includes presentations, lectures, labels, mailing lists, instructions, schedules, invitations, memos, letters, notes, essays, reports, white papers, images, budgets, and work orders among others. Some of these records were not shared extensively across the organization, except with immediate area co-workers or supervisors. For example, the receptionists did not create many records, and the ones that they did create, such as email and phone lists, instructions on how to take messages, and how to use the phone system, were for their exclusive use. All staff members received budgets, bills, reservation confirmations, supply lists, bibliographies, articles, etc. from external sources in relation to the various projects they were handling. These were mostly shared between area employees and with the financial manager. In addition to the letters offering grants that were generated by the grant tracking system,[111] managers and coordinators answered individual requests or sent information through formal letters. Because these letters were not related to a particular project, unless the area assistant or coordinator kept a general correspondence file, the only existing version of these letters remains in the shared directory.

Other work-related record types were part of the communication and consultation with external evaluators and consultants, grant recipients, potential grant recipients, contributors, service providers and the general public. Either as summaries, keywords, lists, plans, titles, reports, or recommendations, all these work records made their way

into the official documentation. For example, Pedro edited all the books published by Aleph whose subjects included historic photography, conservation and restoration, art history, sculpture, and museum policies. From the first publication to the last one, he kept records of all the versions he exchanged with the contributors. As specific projects of the foundation, the editing/publishing of books was proposed and had to be approved by the board of directors through the board meeting minutes and their progress was reported in the annual and quarterly reports.

Natalia's folder contains movie scripts and resumes in Portuguese and in Spanish submitted electronically by grant recipients from one of the foundation's regular training programs in Visual Arts. These were translated into English and sent to the instructors for consideration. These materials made their way into the official documents as a summary of the training program and as lists of participants. Regularly, the foundation commissioned international experts to audit its programs or to review developments in target fields that ranged from libraries, microfilming, mathematics, marine science, chemistry, molecular biology, agricultural science, and physics to preventive conservation and education in elementary and high school. These reports live in the directories of all the staff members involved in the Area of Science and Education, including various working and final versions in English and Spanish. When the final versions were complete, they were appended to the board meeting minutes.

Text fragments

Incomplete pieces of text are ubiquitous inhabitants of the shared directory. Loose notes without titles or dates, letters showing the addressee but without further content or content without the addressee, and lists of emails or mailings with no explicit purpose, are some examples. Other types of text fragments are related to the processes of editing

and exchange that went back and forth between staff members to produce final versions of official documents and books, including versions of paragraphs whose corresponding book or chapter is not stated and paragraphs that would be added to board meeting minutes, appropriation requests, or annual reports. In all these cases, the context of the work is not obvious from the text itself, and only depending on where they were stored in the directory's hierarchy is it possible to determine the project or subject they belong to.

Personal records

The presence and content of personal records in the shared directory can be described as subtle. By this I mean that these records are not extremely personal or revealing. I found school related records such as articles, lectures, class outlines, and rosters belonging to staff members who were students or teachers. There are also records related to the professional associations or interest groups to which some staff members belonged. I also came across resumes, letters with claims to the phone company and to a condominium's administration, reviews of shows, tourist information, invitations and lists of guests to a wedding; images of trips, family members and friends, jokes, and articles of general interest.

ARCHIVES WITHIN THE ARCHIVE

Electronic record-keeping practices were unfailingly related to the paper record-keeping system in some way, except that the processes involved and the values given to each format varied with the person in charge of creating and maintaining the files. Eliana maintained various filing systems: hers, those of the presidency, and her boss's, both electronic and paper formats, for each of which she had different standards. While nobody told her what to keep or discard, as assistant to the president she understood that

the records received and gathered in her area were important and could be requested at any time, so she kept almost everything in paper form.[112] After a period of two to four years however, she discarded memos, letters, and official documents in electronic format from her directory, mainly because she had already printed and filed them or transferred them to her boss's folder in the shared directory. She did not apply this rule to a sub-directory in which she kept records from 1998 and 1999, a period during which she worked with a project in the Arts and Cultural Heritage area.

First in a paper notebook and later in a spreadsheet, during the twenty years that she worked at Aleph she maintained a registry to keep track of the presidency's paper files. The registry included date, topic, the record's ID, and its location in the office file cabinets. When I asked her whether she had devised a naming convention she went to her work-station to look up her folder in the shared drive to see if she had one or not. As we reviewed the files together she became aware that not only did she have a relatively well established naming system, but she had also created one for her boss's directory, which allowed him to find records easily. When my questions made her aware that she had never been conscious of her electronic record-keeping system she said, "It takes awhile to understand the tool that you are using." [113]

Jose, Aleph's president, did not create records, but his awareness of information about everything that was going on was fundamental to comply with his fiduciary role and to represent and defend Aleph's projects before its board and the board of the umbrella organization. To keep him up-to-date Eliana created a directory structure both in the shared directory and on his personal laptop where he could find the final versions of official documents and the progress reports of major projects undertaken in the

organization since 1998. About the contents and convenience of his private digital archive he said,

> *I have an electronic archive that I use, and it is very useful for me because I can't walk in the streets with the paper archive. What I have and I take to all the meetings, things that are interesting to me are in my personal computer, my lap-top; they represent a synthesis of my vision for the organization… I can't trust my own memory regarding the time in which things happened; I can remember what but not when, so when I go to a meeting I have all the annual reports, all the action plans, all the documentation of the umbrella organization, everything. This is the only place where everything is together, in this little machine. My secretary updates this directory once a month. It is very well labeled. If someone asks me about something, I go to my archive, and I can say in what moment, month and year something happened.* [114]

To distribute funding, the Arts and Cultural area called for submissions, received individual requests for projects, and generated foundation-initiated projects whose themes changed according to the needs observed in the field. The work-load in the area was intense, requiring two full time assistants, and temporary staff to produce and supervise the myriad projects that ranged from organizing and producing international conferences on the quality of public television, to funding a rare book conservation lab in a Benedictine monastery. As the area's principal assistant, Natalia was the hub for information from grant recipients, external evaluators, and other assistants which she later piped into the official documents in the form of a summary. Complementing the creative role of her supervisor, the cultural manager, she formalized his drafts and the ideas that he discussed into action plans and appropriation requests that would form part of the official documents of the organization and she also wrote most of the letters that were sent out with his signature.

When she started working in mid 1997, she inherited her predecessor's directory. Over this base she built her own structure using functions (calls for grants, board meetings, etc.), project names, and subjects (conservation, cine and video, museums,

music, etc.) as high level organizational hierarchies. Within these categories, she grouped items by year, by project or by personal name, or did not use further organization, and as a result each unit had its own logic. Her file naming convention choices ranged from very descriptive and long to short and hieroglyphic, whatever helped her remember what the document was about. From time to time she did some house-keeping: re-organized her files and incorporated the files that part-time project assistants and former coordinators in her area had produced for specific projects.

Her virtual folder contains more files than anyone else's, more than double the size of the second biggest directory, whose owner worked in the organization over the same period of time. Besides keeping materials in her extensive digital corpus, she tended to print most of what she received in relation to projects and to include it in its respective project file. About the role of the paper and the electronic files in her work she said,

> *The paper files, being those subject or project files, gave me a sense of organization, of things in place, of unity that the electronic files did not. In the shared directory you had to open one document after the other, and the sequence of the project or the theme is not as easy to reconstruct as when you flip through the paper. The project file was something that everybody understood, but the individual electronic folders were not. For example, I could not access other people's emails to see their communications with a beneficiary, plus I never organized my email mailboxes, but everything was on the project file. The problem was finding an electronic version of a paper record, it was impossible! So, things were not managed from the computer but from three different systems.*[115]

The process of tracking projects in the Social Welfare area involved complicated budgets and scheduling adjustments. To keep track of all the details, the area manager, Celia, classified and weeded her email and other electronic records regularly, discarding information that she considered ephemeral and keeping what she considered significant. Her rule of thumb to keep or discard was "When in doubt, I keep."[116] Her first level file organization scheme consisted of subjects within which she arranged the files, according

to the currency of the issue, in folders named “old” and “in use.” As soon as beneficiaries sent in progress reports or correspondence in electronic format, she added them to the appropriate electronic folder. This practice allowed her to have almost the same representation of the project in electronic format and in paper; it was even more complete in electronic format because she produced project analysis documentation that she normally did not print out. Like Natalia, she named her files using keywords, institutional names, project numbers, or whatever made her remember what the record was about, and like Eliana, she went to her computer to explain about her arrangement because she could not remember the details off the top of her head. About her patterns of arrangement she commented,

> *I changed my system many times. The archive that I made was not always the same, but since I did not have much time unfortunately all the configurations are incomplete, because I implemented something that for me was an improvement, but I did not have time to update everything that was on the previous form. So I only placed the active documentation in the new format but not what before had been managed in a different way.*[117]

In the area of Education and Science, established grant programs suited the more structured profile and the competitive tradition of the communities of hard scientists, doctoral and post-doctoral candidates, and honor students. This structure allowed for easy tracking of the beneficiaries’ progress through the database, but it did not reduce the amount of work involved in launching each call for proposals, nor did it make the distribution of the electronic records more organized. Diana started collaborating at Aleph in 1993 as a free-lance to coordinate specific calls for grants, but none of the records that she produced was stored in the shared directory, as she would write documents at home and turn in printed copies. In 1997 she was assigned a computer at Aleph—the same one that was used by all the free-lance coordinators—and started to spend more time at the office, but the records that she produced at that time remained in

the directory of some other area employee or in the paper file. In 1999 she was assigned an office, her own computer, and a folder on the shared server, and she became a full time employee in 2002. The files in her directory, dating from 1999 to 2005, show that only 11% of the records belong to the period 1999-2002.

Because of his roles as executive director, manager of the Science and Education area and editor of Aleph's publications, Pedro's directory contains records related to all three functions. He separated records according to their format or function (images, Power Point presentations, texts, and documents that were posted to the web) and separated current from old. As a naming convention for files and folders, he maintained the 8 character convention inherited from the DOS era. Traces of his record-keeping system, such as the short naming convention and the division of old and current files, are present in the directories of two of his closest collaborators, but only diffusely and randomly. As in the rest of the areas, the only record-keeping pattern was that there was none. The electronic record-keeping system was not influenced by the managerial style of the area or by the community of practice that it serviced, but by the needs, imagination, and style of each individual person.

Traditions and Transitions

The dynamics of use of the databases, the paper records, and the electronic records reflect how these systems coexisted, at times overlapping and other times complementing each other. It also reflects the tensions between the strong paper tradition and the recent and somewhat mistrusted electronic one. The record-making and -keeping practices mentioned in this section are post 1997 and were followed until the institution

closed in 2006. The recollections gathered during the interviews were collated with observations of the content and arrangement of records present in the shared directory.

Local legal particulars require philanthropic organizations to file every year the audited annual balance with the General Inspectorate of Justice. As of 2005, the document had to be filed on paper and signed by the foundation officials. To produce these records, the financial manager, Ruben, exported the yearly data from the financial database onto a spreadsheet in which he had pre-inserted formulae. The results were transferred to a table in a Word document which was printed and signed for official presentation. A second paper version was copied to the legal accounting book, and a third one was kept on file in the manager's office and used for reference and auditing purposes. Given that the raw information was in the financial system and he could reproduce it any time, Ruben did not keep the electronic Word document or the spreadsheets, which were templates that he populated every year with different data.

From 1998 and until the last project was closed in December of 2005, Elena tracked the projects' progress and wrote quarterly reports in English for the umbrella organization. As she read each report sent by a grant recipient she entered a summary in Spanish in the grant tracking database and copy pasted it onto a Word document that would later serve as the basis for the quarterly reports. In this way, she did not have to go back to the database system or to the project file to check the progress of each individual beneficiary. To create each new quarterly report, she used the one that she had generated three months before as a template and saved it with a different name. Since many of the same scholarships were active between and throughout years, many headers used in the previous report remained the same, and only the content had to be modified. She devised a different file naming convention for the summary document and for the finalized

quarterly report and stored both in her directory, one in Spanish and another in English. She did not delete any of the electronic versions, but printed and filed the final English reports on paper.

These examples show different interactions between the databases, the electronic records, and the paper ones, as well as differences in the values given to the paper and electronic records: Elena kept everything electronic, and Ruben did not. In the latter case, the difference in appreciation was informed by the evidentiary weight of the paper document for the General Inspectorate of Justice and the external auditors. Also, the affordances of the different systems played a role. Combined with the spreadsheet, the financial database allowed manipulating the data more efficiently. The grant tracking system, on the other hand, did not help in reporting multiple projects, and Elena had to create her own short-cut by copy pasting progress summaries to a Word document.

The uses of the project file also constitute an example of tensions. Project assistants and coordinators were in charge of creating and updating the project files following the unwritten rules initiated by Renee. In addition to what constituted the official documentation—namely the grant application form, a copy of the approval letter, the signed letter of acceptance, and the signed general conditions form and background materials required in the call for grants—each file could also contain progress reports, communications (letters, faxes and printed emails), notes, internal memos, evaluators' comments, final products, and other records deemed important by the assistants. Videotapes, CDs, and DVDs submitted as part of the application were stored separately in the foundation's library and were controlled through a database that linked the item to its project number. While the project was active, the file was kept close to the assistants' work-spaces. Especially those projects initiated by the foundation could occupy several

heavy dossiers. Upon a project's completion, the file or files were housed in compact mobile shelving in a room interchangeably called "basement" or "archive" and seldom accessed.

In the early years, active project files were heavily used by managers and assistants and circulated among the members of the Board for review during their monthly meetings. Once the volume of grants increased, the circulation of project files was replaced by a document listing the projects' basic information and the recommendations issued by the area managers. As for their day to day use, the summary of the project in the grant tracking database decreased the need to access the entire file. And yet, in order for payments to be sent to the recipients the project file had to be presented to the area managers first, the financial manager second, and finally to the executive director or to the president, who were legally authorized to sign checks. During payment days, project files would pile up in the offices.

When I asked the financial manager why they needed the paper file when they had the progress information and the payments scheduled in the grant tracking database, he told me that it was customary to attach all the background documentation, as he could not go to the director or to the president with a bare check to sign, a practice that suggests a ceremonial use of the paper records and indicates the authority that they carried. But it also had a practical side. The financial manager wrote annotations on the letters of offer regarding adjustments in the dollar/peso exchange rate value that affected the payment commitment and those notes were reviewed by the authorities at the moment of signing the checks.

WHAT IS THE ARCHIVE?

User's point of view

Most interviewees agreed that the paper files constitute Aleph's official archive. The president of the foundation explained how all the actions and activities of the foundation, from planning to deciding, to acting and evaluating, were more or less summarized, in the different official paper documents, the final abstraction being the annual reports. He said that the electronic records were never meant as archives but as instruments in aid of writing, copying, sharing, editing, rewriting, and distributing. Staff members in charge of project coordination also referred to the paper project files as the most complete rendition of the organization's accomplishments and as evidence of their individual work. The latter was substantiated as they printed and filed all their electronic communications with beneficiaries and evaluators as well as the memos that they presented to the administrators in relation to any particular case. The project file containing the complete story of the grant transaction backed up their decisions and provided answers to questions made by the administration. However they also explained that they could not have managed without any of the three systems in place, since each of them played a different role in their daily work.

Celia, the social manager, summarized the different functions played by the three systems and their limitations. The grant tracking and financial databases had specific administrative functions, such as generating a unique number for each project and the letter of grant approval as well as allowing payment authorizations and verification of payment status. However, its progress tracking and reporting features were limited. The system did not provide a complete panorama of the project's progress, mostly due to its interface which only allowed viewing one progress report at a time. In addition, the space

available for reporting was insufficient and the scheduling function (in which dates for reporting and payment were set) was rigid and did not allow changes. About the shared directory on the networked server, Celia explained that it allowed exchanging records and manipulating data and text to create new documents. However, depending on how each person arranged, named, and updated their files, it could be useful or absolutely useless. Finally, the paper file contained the legally binding documents and included materials sent by the grant recipients that were not elsewhere. This testimony confirms the generalized conception that the paper records were considered the official archive in the organization.

Electronic records

And yet, the narrative of the archive's formation process reveals a less black and white picture of the roles and uses of the electronic records in the organization. We learn that even though these were difficult to find in the shared directory, many staff members made daily decisions on how to organize and name them, and that they continually tried to improve their arrangement. Also, and despite the fact that they could not use them efficiently, the general tendency among the staff was to maintain their electronic records over time. Moreover, while some staff members deleted electronic records that they did not consider valuable at a specific juncture (time sensitive materials, notes, general communications, etc.), they kept those that they considered important, either to document their work or because they valued the project or the event that generated them. Additionally, some people created and maintained electronic records that they never printed, and some transferred all of them to their laptops so they could carry them everywhere.

To all, the contents of their virtual folder constituted their personal work-collection, and across the organization the only consistent behavior was that each employee had a unique way of managing it, including strong ideas about the adequate etiquette for sharing the records with co-workers. Despite their sense of property and privacy, upon leaving the organization most staff members left their folders intact, with records belonging to all their working years and some personal ones. This unspoken practice was also followed by IT consultants and systems administrators. Neither aware of any kind of IT guidelines nor urged to create them, they too left vestiges of their work on the networked server in the form of software, scripts, and systems that they created and used, and did not delete those left by others before them. The natural archive also shows that the boundaries between private and public are clear when they are enforced but otherwise they exist in a blurry mode.

These behaviors can be interpreted in different ways: that the staff members wanted to leave an evidence of their activities; that they considered that these records and objects could be of use to others; or that they perceived that they were property of the organization. Given the freedom to do as they pleased, in some cases instinctively, in others without much consideration, and in others purposely, staff members naturally chose not to let go of the electronic products of their daily work and their personal lives. This behavior suggests that the electronic records were considered of current and future value by the staff members. It is only when they are contrasted against the paper records that the concept of choosing between one and the other emerges.

Paper and electronic records

The structure of the paper files and that of the electronic records in the shared directory differ.[118] The inventory of the paper portion of the archive, created during the

process of arrangement and re-housing, shows that at the group level the paper files map the formal organizational chart, simply because the records existed within the office areas defined by the organization's structure.[119] At the series and sub-series level the paper files contain paper records sent to the organization. Differently, the first level structure of the shared directory shows the staff members' virtual folders with their initials and their personal electronic records collections. And, as stated by Natalia during her interview, "no efficient way was ever devised to connect the electronic record with its copy in the paper file."[120] In general, and while some people included records sent to them in electronic format in their virtual folder, the exchange between the inside and the outside of the organization is present in the paper files, within which the project file is the one that better reflects the impact of the foundation's projects through the reports of the beneficiaries.[121]

At the series and sub-series level the structure of the paper and electronic files are also different. In addition to the project file, the rest of the paper files are area-centralized, containing series with records generated and gathered by the staff members in a given area (bear in mind that areas at Aleph were small in number of employees and mostly the assistants filed their supervisor's records), while the virtual files contain what each staff member decided to keep individually. So, for example, the fact that some members worked for more than one area is reflected only in the virtual directory.

Most notably, it is the electronic presence of fragments, versions, and the repetition of official records that fails to match between the paper files and the shared directory, because in the former only final versions are included. Also the distribution of these records is different in both systems. For example, the board meeting minutes in the paper file—kept in chronological order—is a complete series. Copies are also placed

individually in the project files to assert the meeting in which the project was approved. In the shared directory, on the other hand, these records are irregularly distributed throughout the staff members' folders. The paper files present a single functional provenance for most series and sub-series, but in the shared directory, records have shared provenance across staff members and their corresponding functions. Finally, paper and electronic records series complement each other in providing a complete picture of the foundation. To give one example, the receptionists are not reflected in the paper file, but they are in the shared directory. Conversely, the financial function is better reflected in the paper files because—as he stated in the interview—the financial manager printed and filed the majority of the records that he generated.

This research also surfaces aspects of the transition from conducting transactions through signed and sealed paper documents to conducting them through electronic systems and records. At the same time that the organization was actively implementing IT solutions, the strong cultural roots of the paper system did not allow it to take full advantage of the tool that had just been acquired. In the narrative, the decisions to incorporate new IT tools come through as moments of bliss, drive, and opportunity; but in each advance there is a component of attachment to old habits, of resistance to change, and of ambiguity derived from having to deal with the hybrid of paper and electronic records.[122] During the institution's active life the paper and the electronic systems were never fully integrated and each staff member found a unique way of putting the pieces together to pursue his or her work.

Mixed with official and work documents in the share directory, records sent by people outside the foundation and personal records provide information of the broader context in which the organization functioned. The former shows the interests and

concerns of those who interacted with the organization, the latter points to a time in which many people did not own computers and used the ones at work for personal matters. It also shows how work and life are intertwined and can only be separated through restrictive regulations such as for example those that started appearing in the early 2000s in relation to the use of email.[123] The way in which the staff members used and maintained electronic records and systems on the server suggests strongly that electronic record-making and record-keeping were the processes behind their actions, the tacit routines that led to transactions that were paper bound, or better said, that became explicit when they were printed, signed, sent, or received on paper media. That at the very end, the organization implemented a virtual grant approval and offer system indicates that it was moving towards a more active use of the electronic environment and less use of the paper records.

An archival perspective

The question of what constitutes Aleph's archive should also be considered from a formal archival perspective. Just as when they were in use, the functions and values of the three records systems (paper and electronic records and databases) cannot be separated. For example, while each project file ideally presents a complete rendition of the activities undertaken by a beneficiary, the fact that the project file may be incomplete due to weeding and flood damage, points to the grant tracking database system as an indispensable source to complete—even if not in detail—the information about all of the projects undertaken by the foundation. Furthermore, since confidentiality commitments prevent public access to the project files, the value of the database systems as a repository of redacted information rises. From a legal perspective, current practices indicate that the tendency of judges is to request electronic records and the databases as evidence.[124] Since

over the course of the institution's active tenure the records had not been discarded, keeping them is a serious concern from this point of view.

The implication of the presence of personal records in the archive poses the question of whether they should be permanently retained, restricted from public access (if it comes to it), or discarded. For many staff members this was their first time using computers, they did not have one at home, so they used the one at work for personal errands. Others had a computer at home and still used their time at work to generate these records and use the shared directory to store them. While the use of work time to conduct private affairs is not new, it would not be reflected if the records had been managed in a controlled record-keeping environment. Indeed, there are no such types of records in the paper files. Considering that the archive was created during a period in which concerns with electronic records and privacy were yet to be understood and in the context of a culture that is less aware of boundaries between private and professional, the presence of these materials reflects the habits of the records creators. In this case I decided that these records are components of the natural archive, a decision facilitated by the fact that the future access to the archive is uncertain. But I also acknowledge the challenges that would arise if these records were to be made public.

Aspects related to provenance, versions, paragraphs, and gaps and complement in the archive are clarified through the narrative. While the paper file shows the final versions of records and in the project file in relation to the project to which they are related, the process of making the record is best reflected in the electronic archive. The records-creation workflow shows why parts of one record are distributed amongst the staff members' folders, explaining its shared provenance and highlighting the importance of location—understood as file path—to establishing who contributed what and when in

the making of the official version. In turn, the evidence of work processes ratifies the authenticity and the integrity of the component pieces and of the final record.

Aleph's archive

As a whole, the narrative of the archive's formation process allowed concluding that the archive is one, composed of paper files and the natural electronic archive, and that separating the components or choosing one over the other would be arbitrary.

Material culture perspective

Around the shared directory, the vestiges of old applications, scripts, and drivers provide clues to how digital archives were created and used at a time in which there were no conceptions of what an electronic record-keeping system ought to be, nor what electronic records meant to individuals and to society. Removing this digital evidence will restrict the possibility of learning about the way in which computers were introduced at the time of massive adoption of information technologies and how the values, uses, and perceptions of paper and electronic records changed over a period of twenty years.

The natural archive concept

The study also revealed the unique characteristics of the electronic portion of the archive. Considering the networked server akin to an archaeological site allowed a systematic approach to analyzing the phenomena in ways that might not have been possible if applications, records, and systems found "in the site" were dealt with separately. I concluded that the manner in which records and tools had been kept on this server could not easily be ascribed to digital archiving models currently discussed in the literature. These models focus more on the creation of "sound" electronic records, the design of electronic record-keeping systems, and on institutional repository archiving models than on the way in which digital archives are actually created.[125]

As a result, the concept of a "natural electronic archive" started to shape. The term refers to an expression used in the Spanish language to refer to activities that people do without much afterthought, without forcing themselves, encompassing actions that are instinctive, in which intentionality is blurry. It builds partly on the definition of "natural collections" proposed by Phillip Cronenwett to describe collections of literary manuscripts that are not fragmented as they leave the hands of their creator.[126] It also reflects Hillary Jenkinson's characterization of archives as natural accumulations as a consequence of the conduct of affairs.[127] The natural archive (that is further explored throughout the dissertation and defined) as a unit of analysis, allowed looking at the contents of the networked server under the idea that they all can be read as records, that they are all evidence of the past, and that preserving them allows revisiting the organization from multiple perspectives.

"EXCAVATING" THE ELECTRONIC RECORDS

The narrative of the archive's formation process presented above shows that Aleph's archive is hybrid, composed by paper and electronic records and systems. The structure of the shared directory and the technical properties of its files anchor the plausibility of the narrative. And, similar to an archaeological site, the rest of the digital objects surrounding the shared directory provide context and legitimate the electronic records within. However, the quantity and chaotic structure of these records slows access and impedes their analysis as a way of answering the question of whether they represent the institution that created them. For this reason, I designed a digital appraisal method that combines archival principles, text mining, social network analysis, and visualization. The method, described in Part III, takes advantage of the atomic nature of digital records,

creates structure out of the chaos, and extracts evidence about the organization from its electronic text records.

PART III: PRESENT

Mining the Natural Archive

Appraisal is the process through which archivists determine which records will be retained for the long term in the archive. Having established the contents of the networked server as a natural archive, the next step in the case study was finding out whether the records in the shared directory represent the activities and roles of those that created and maintained them. The next section provides an overview of the state of electronic records appraisal and a platform to introduce the digital appraisal method designed for the text records of a natural archive.

APPRAISAL REVISITED

That archives provide evidence of what society does, thinks, and believes is the basis of archival discourse.[128] The heart of archival appraisal is to determine how to select and thus to identify records of enduring value as they relate to societal actions, ideas, and beliefs.[129] Over the years, the criteria and methods used to appraise archives have elicited intense debate. Adding complexity to the discussions, the emergence of electronic records, their nature that allows transformation and change, their strong relationship to the tools with which they are created, the aura of vulnerability and risk that surrounds them, and—as in this case—the lack of control with which they are created and maintained challenge the ability to establish whether they provide evidence of the people and organization that created them.

As a response to the problem of managing large amounts of paper records, in the 1950s Theodore Schellenberg and colleagues at the National Archives in the United

States established that the essential task was to identify primary values—financial, legal, fiscal—and secondary values—evidential and informational—in records, and establish records retention schedules accordingly. Among the secondary values which are identified as archival values, the main criteria to determine records' long term retention is their evidential value, understood as information that renders an accurate representation of the organization or person that created them.[130] In the United States, at the government level this conceptual framework is still used to appraise electronic records.[131]

By the mid 1980s the confluence of the emergence of electronic records and the perception that archives were not holding the documentation demanded by society, embarked archivists in a review of the state of the profession.[132] In Canada, archivist Hugh Taylor addressed these issues with an intellectual freedom that prompted a brainstorm of ideas and suggested the opportunity to insert archival activity at the center of social and historical inquiry.[133] Under the theoretical umbrella of postmodernism, many archivists acknowledged their role as shapers of the record.

In this juncture, archivists saw the opportunity to develop new appraisal theories. Searching for new ways of capturing records of permanent value, novel appraisal methods such as documentation strategies and functional analysis emerged.[134] As new forms of record-making and more flexible organizational forms developed in the organizational world, these methods morphed into macro-appraisal and network appraisal. As a top-down approach to capture permanent records on the basis of social analysis of organizational functions and structure, macro-appraisal promised to reflect the impact of government on the public.[135] On the other hand, for smaller and ever changing organizations, network appraisal argued that capturing records of permanent value should focus on analyzing work processes and changing relationships among group members.[136]

More conservative lines of archival thinking maintained that archival science should distance itself from philosophical tendencies and emergent trends. From this perspective, to constitute evidence records of any format have to be authentic, that is legally and administratively accountable and historically trustworthy.[137] But in the fuzzy electronic environment, to produce such records archivists would have to participate in the design of electronic record-keeping systems and embed archival functions within them.[138] Throughout the 1990s, research projects such as "The Functional Requirements for Evidence in Recordkeeping"[139] at the University of Pittsburg and the "Preservation of the Integrity of Electronic Records Project"[140] at the University of British Columbia (UBC) applied different traditional archival ideas and disciplines—the former legal and best-practices conceptions of records and the latter diplomatics—to determine the requirements that record-keeping systems should have to produce authentic and reliable records of evidential value. The document "Design Criteria Standard for Electronic Records Management Software Applications,"[141]an outcome of the UBC research project, and its subsequent implementation and certification processes in electronic information systems constitutes an example of such approach.

Gradually, authenticity as informed by diplomatics became understood not only as a fundamental function of electronic record-keeping systems but as criteria to appraise electronic records.[142] InterPARES I, an international research project in electronic records that lasted from 1999 to 2001 specifically emphasizes this approach. One of the projects' outcomes is a deductive diplomatic model that specifies a series of requirements that electronic records have to meet in order to be considered authentic and therefore eligible to be considered of permanent value.[143]

This strict view of electronic records has been challenged. In his essay "Afterglow: Conceptions of Record and Evidence," Brien Brothman explores the relationships between evidence, truth, and electronic records in the prevalent "strong sense" conception of records described above.[144] He wonders whether archivists are aware of the consequences of this view and discusses the validity of "weaker sense" conceptions of records that apply to private documents and records not created under legal regulations. Furthermore, he shows how views that separate evidence from authorship, use, or particular record-keeping systems are the ones that prevail among historians and record users. Firmly, Brothman points out that archivists' emphasis on creation of perfect record-keeping systems is opportunistic and linked to the rise in status of information management and corporate interests.

Interestingly, the results of InterPARES II (the continuation of InterPARES I which lasted until 2006) somewhat supports Brothman's concept of weak and strong sense records. As a result of analyzing case studies of electronic records against the authenticity standards devised during InterPARES I, the conclusion was that almost no electronic records could be reported as authentic, and that while useful, diplomatics falls short of handling the diversity of systems and electronic record types available. From focusing on individual records, InterPARES II evolved to assessing the integrity of the information systems and processes from which records emerge by establishing different levels of authenticity that accrue to records as a result of how these systems and processes are set up and used. [145]

Despite the use of functions, structures, and work processes as objects of analysis, or of considering lower and higher levels of authenticity or weak and strong sense records, appraisal is always ambiguous. Submitting electronic records to strict models of

authenticity requirements may lead to discarding all or most of them, but considering the particular context of the information system from which they derive can save them. Judging electronic records from the perspective of the organizational structure and functions may lead to keeping them all or only final versions. But it can also lead to disposing of many depending on whose perspective is favored; whether the electronic records are perceived as duplicates or references by their creators or chaotic and unapproachable by the archivist. In all cases, the value of records still emerges from assumptions made about people, organizations, functions, or the particulars of the information systems from which they come, all of which vary significantly from one case to the next, making it difficult to apply standards to the appraisal process. Moreover, none of these appraisal methods answer the question of whether the evidence provided by electronic records is more, less, or equal to that of paper ones.

It is worthwhile to revisit the thinking of English archivist Hilary Jenkinson, and reflect on the warnings that he issued during the first half of the 20th century about the risks involved in imposing values on records and of practicing appraisal as selection.[146] Jenkinson considered records as impartial by-products of transactions and only the archive's creators, in the context of work needs and processes, could decide which records had to be retained for the long term. To him, the role of archivists is to take charge the archive as it comes from its creators, and not disturb it by making timeless judgments and selection. My interpretation of Jenkinson is that when he spoke about the impartiality of archives he did not mean to imply that record creators are saints[147]or that all transactions carry some sort of essential truth and records cannot be biased.[148] To me he meant that the archive, preserved as it was created and used allows that whatever truth or bias exists in the records will surface as long as archivists do not interfere between the

archive and the eventual secondary users by selecting or weeding. Archival appraisal has traversed from modernism to postmodernism without resolution, all of which speaks of the need to find other ways of appraising archives and establishing evidence in electronic records disassociated from both the opinions of the archivists and the intentions of the creators.

INTRODUCING AN INDUCTIVE DIGITAL APPRAISAL METHOD

To answer the second research question and investigate what kind of evidence about the organization is provided by the records of Aleph's natural archive, I devised an appraisal method that uses text mining, social network analysis, visualization, and qualitative interviews. I use text mining methods to explore whether the similarities among 16,000 text records created and co-created by staff members during ten years can be used to infer work relations, functions, roles, and organizational changes over time. Results are interpreted with social network analysis theory and in the context of the narrative of the archive's formation process. For validation purposes the results are contrasted with the accounts of staff members about the staff member with whom, when, and for what; by analyzing records' contents; and through analysis of statistical distributions.

The use of automated corpus analysis methods combined with visualizations and interpreted via social network analysis is a trend in the areas of Human Computer Interaction (HCI) and Information Retrieval (IR), mostly geared to email communications.[149] Adam Perer and his team used Ben Schneiderman's email archive to explore patterns and changes between email correspondents over time based on categories of relationships defined by the archive's owner.[150] To make sense of the Enron

email corpus, Jeffrey Heer used data mining techniques on the email headers to establish various forms of relationships and linked those to the possibility of visualizing the texts written by the correspondents.[151] In the project eArchivarious, Anton Leusky, Douglas Oard, and Rahul Bhagat combine email headers and email content analysis to identify relationships and information flow between people and use these as a starting point for information retrieval.[152] The area of Text Analysis also uses email and visualization; with the Mandala Browser Stefan Sinclair and Stan Ruecker offer an interface to facilitate content analysis of XML tagged emails.[153]

The digital appraisal method that I present differs in principle with the projects described above because it has the archival goal of determining evidence in the natural archive. Methodologically it also has differences; it does not involve email but more diverse documentary forms which are neither tagged nor classified. The method is rooted in concerns expressed by Peter Botticelli in his study of networked organizations about the need to document dynamics and changes in organizations.[154] It explores the meaning of evidence and the archival bond, the latter understood as the "network of relationships between records," in an ambiguous environment.[155] In this research, archival bond is represented as the similarity, or levels of similarity, between texts written, co-authored and/or shared between staff members over the organization's network. Its overall philosophy is informed by Jenkinson's idea of records creators as selectors of their own records.[156]

To design this method I first conducted a proof of concept during which I learned the scope and limitations of the open source software tools that were available for me to use, the computing resources available, and the way in which the text files had to be pre-processed to obtain accurate representations of the dynamics of the organization. Because

of the nature of the tools used, the appraisal method focuses on the text records stored in the shared directory on the networked server. It does not include the spreadsheets, images, or presentations also present in that storage configuration. The method uses various software tools, some proprietary, some open source, and others especially developed for this research. Both for the proof of concept and to conduct the appraisal method I used copies of the text files found in the shared directory. Below I offer an outline of the main components, steps, and software tools that comprise the appraisal method to be discussed subsequently:[157]

I. Text Mining
    - Pre-processing of texts
        - Constructing yearly samples
            - *File management software – File Boss*
        - Transformation of text records to ASCII text
            - *File transformation software – File Merlin*
        - Sort documents by language to extract Spanish documents.
            - *Daniel Zeman's language sorter in UNIX*
    - Application of Vector Space Model
        - Creation of bag of words representation
            - *Rainbow modified to work with Spanish*
        - Calculation of term frequency-inverse document frequency (Tf-idf) for each document
        - Calculation of cosine similarity measures

        - Calculation of averages and relative averages of cosine similarities over the corpus
            - *Software specifically developed for this research*

II. Visualization
    - Organization level view
        - *Social network analysis software – UCINET*
    - Staff member centered view
        - *Visualization created by the Texas Advanced Computing Center (TACC) Scientific Visualization Team*

III. Research Validation
    - Qualitative interviews
        - 14 staff members and the systems administrators
    - Historical narrative of the archive formation process
        - Based on staff interviews
        - Based on data from the metadata timeline
    - Re-interviewing or member checking
    - Statistical distribution analysis based on Gini curve
    - Observation of contents of the records included in any given relationship.

TEXT MINING PROCESS

Formal definitions agree that the goal of text mining is to discover knowledge that was unknown in corpora of unstructured electronic text.[158] Text mining comprises

various algorithms used across computational disciplines such as Information Retrieval (IR), Computational Linguistics, Natural Language Processing, and Machine Learning. Depending on the research goal, different algorithms are combined for different purposes. In this appraisal method, text mining is used to resurface knowledge from the past which in the present constitutes new knowledge. Specifically, I borrow methods from the area of IR such as the vector space model and cosine similarity calculations, which will be defined as I explain the process in detail.

Pre-processing

Because computational tools are still not capable of dealing with raw natural text, extensive preparation is needed to create an organized work-flow and to normalize the electronic records before they can be mined. The pre-processing steps used to conduct this appraisal method are described in order in the following sections.

Constructing yearly samples

To model the relationships of staff members over time based on the texts that they wrote and gathered in Spanish, I built ten file sets, each containing the text records produced by each staff member who had a virtual folder in the shared drive over the period of a calendar year that runs from January 1st to December 31st of the years 1996 up to 2005.[159] Using FileBoss™ file management software I sorted the files by format and last modified dates, and built file sets with copies of the .doc files from the directories and sub-directories in the shared directory. Each yearly sample contains a different number of documents. Figure 2 shows the directory structure and naming scheme devised for the yearly sets.

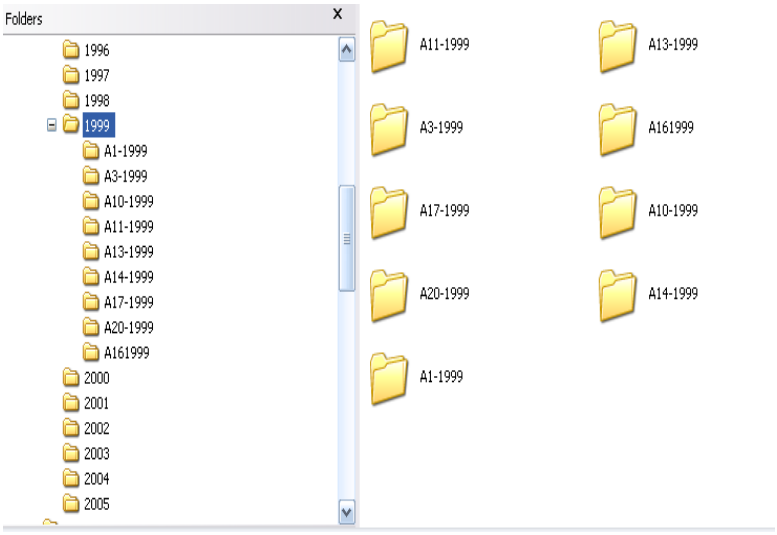


Figure 2: Directory structure and naming convention of the yearly sets.

Each of the sub-folders contains the text records that each staff member created and gathered during the year.

During the proof of concept work I faced processing problems due to the length and the spaces between words in some of the original file names. I used FileBoss™ to rename all the files in the new sets and kept a record of the correspondence between the .doc file and the renamed .txt copy.[160] The file naming convention is: [staff member's code] [last digits of the corresponding year] [record number in the sub-set] example: A1990001.txt.

From .doc to .txt

To be processed by the text mining software, proprietary word-processing encoding was transformed into ASCII text. Text normalization is necessary so that the tokenizer, an algorithm in the text mining software that transforms a string into individual terms, does not confuse word delimiters with other character types due to

differences in the character encoding of documents created with different word processors. To convert the records to ASCII, I used File Merlin™, a file conversion program that supports a variety of formats including Word for DOS versions present in the shared directory up to the latest Word for Windows 2007. [161] The text sets were transferred to a UNIX server to continue with the language sorting and text mining steps.

Language recognizer

The shared directory contains documents in Spanish, English, Portuguese and French, the majority of which are in Spanish, followed by English. Most of the English texts are versions of documents in Spanish, and although there are specific records in English that do not have a Spanish version and vice versa, much of the same information is present in both languages. Since Spanish is the dominant language present and is spread uniformly across the staff members' directories, I decided to use only the texts in Spanish to conduct the appraisal. For this I needed to find a tool to sort records by language.

The search was not easy as the language recognizers that I found online were expensive, part of larger applications, or only worked with one file at a time. I placed a question about language recognizers to the Corpora List,[162] and Daniel Zeman, a scholar in the Institute of Formal and Applied Linguistics in the Czech Republic, sent me a language recognizer and sorter that he created in Perl. The program works with batches of texts, and Daniel went out of his way to modify it so that it would work with the directory structure that I needed to create for the text mining process. He also sent me detailed explanations on how to train the program in the languages that I needed to sort out.

Final sets

The number of staff members and of files included in each file set varies. Table 1 shows the configuration of the final ten yearly sets.

Table 1: Conformation of the yearly sets.

Year	**Number of records**	**Staff members**
1996	544	6
1997	719	8
1998	1199	13
1999	1374	14
2000	1595	13
2001	2181	14
2002	2681	15
2003	3727	16
2004	1677	17
2005	707	9

Vector space model [163]

A vector space model is a mathematical representation of a given corpus in which each document is a vector in a multidimensional space. This model allows calculating similarities—understood as similar words—between the documents included in the corpus. Similarity is based on similarity between the words used in each of the documents that are being compared. In this appraisal method I built a vector space model

from each yearly set. The process starts by transforming the documents from each yearly set into a "bag of words representation" which contains all the words included in the set. For this, the documents in the set are tokenized and statistics of the number of times in which each word appears in each document (term frequency) are generated. In this way, the entire vocabulary in the set is represented as a vector space whose dimension is equal to the number N of unique words that together constitute the corpus vocabulary of each yearly set. In this N dimensional vector space, an individual document is represented as a vector with N components. Each of the N vectors' component is a word with a measure of the frequency with which that word is used in the document. Thus, each document is scored for its use of words from the whole vocabulary. In the vector space model, documents with similar word frequencies are located close to each other and then measured with the cosine similarity distance calculation.[164] Figure 3 below shows a representation of documents in a vector space model. In the first representation, each document is represented as a series of term frequencies; note the way in which the provenance of the document is preserved in the file path shown in bold. In the diagram, each of the documents are represented as vectors (when looking at the diagram it has to be considered that it is a two dimensional representation and the vector space model is a multidimensional representation).[165]

/1991/A1/A1990001.txt/A1,experto 1,programación 1,actividades 1,culturales 2,abogado 1,artista 1,visual 1,ganador 1,primer 2,premio 2,nacional 1,municipal 1,arte 2,dirige 1,programa 1,cultural 1,incluye 1,artes 1,conservación 1,….

/1991/A1/A1990002.txt/A1,experto 0,programación 0,actividades 1,culturales 1,abogado 0,artista 10,visual 3,ganador 0,primer 3,premio 0,nacional 13,municipal 1,arte 24,dirige 0,programa 4,cultural 8 ,incluye 0,artes 27,conservación 17,….

/1991/A1/A1990003.txt/A1,experto 0,programación 0,actividades 1,culturales 1,abogado 0,artista 0,visual 0,ganador 0,primer 0,premio 0,nacional 0,municipal 0,arte 0,dirige 0,programa 0,cultural 0, incluye 0,artes 0,conservación 8,….

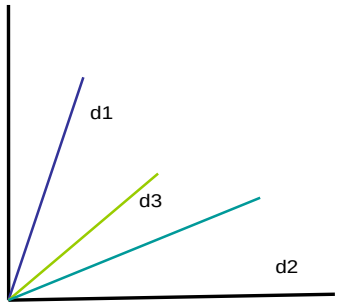


Figure 3: Vector space representation of three documents

During this step stop-words—common words like articles, prepositions, connectors, numbers, and even common verbs—are removed from the text sets based on the assumption that they are so ubiquitous that they are not useful as pointers to distinctive features in texts. This step also reduces the vocabulary significantly, and

therefore the computational resources needed to process the texts. In each set I used a Spanish stop-word list to indicate which words had to be removed.

Stemming is the process through which words are reduced to their stem or root. It may or may not be part of building the vector space model. For example, in a corpus containing the words "house" and "housing" both will be considered instances of the word house. Because it reduces the length of words, the computing resources needed to process a stemmed corpus are considerably less than those needed to process the same non-stemmed text. I was interested in introducing a stemmer in the appraisal method to test the difference in results obtained between the stemmed and non-stemmed texts and in relation to obtaining a faithful representation of the records. Also, I was concerned with reducing the computing resources needed to process the sets, as I had experienced processing memory shortages during the proof of concept. After experimenting with stemmed and non-stemmed file sets I decided not to use this feature. The reasons are explained in the section *Relationships based on texts*.

Text mining software: Rainbow phase

To tokenize the texts and build the bag of words representation, I used Rainbow, an open source text analysis and classification program written in C++.[166] Because it is a classifier, Rainbow works well with the sub-directories or staff member's classes included within the yearly sets (See Figure 2). Some aspects of Rainbow were modified for the purposes of this study. The tokenizer's code was changed to recognize Spanish language characters such as accented letters as specified by the ISO/IEC 8859-1 Latin alphabet 1 standard.[167] A Spanish stemmer [168] was added to be used as an option and a Spanish stop-word list was included; these additions had to be programmed within Rainbow to be used from the command line. These modifications were introduced into

Rainbow by Dr. Hai Bi, who, when I first contacted him to work as a programmer on this project in 2005, was a doctoral student in the School of Electronic Engineering at the University of Texas at Austin. He also coded the software used to complete the next phases of the appraisal method. As a result of processing the ten file sets with Rainbow, I obtained ten matrices as .txt files, each containing the document path and name, and the absolute term frequencies of all the words in the vocabulary per document in the set after stop-word removal.

Words must be weighted and not counted [169]

All the words in the vector space model are related to all the documents by way of their absolute frequency; that is the number of times in which they exist in each one of the documents in the set. There are other ways in which word frequencies are calculated, and each can render a different representation of the same document.[170] For example, when word frequencies are normalized—to yield relative frequencies—the differences in length between the documents involved in the set are balanced. Another normalization alternative is the term frequency-inverse document frequency approach (Tf-idf). Tf-idf considers the length of the document in which a word appears, whether the word is rare or common in relation to the document, and whether it is rare or common in relation to all the documents involved in the set. [171]

With this calculation, the importance of a word increases proportionally to the number of times it appears in the document, but it decreases if the word is very common throughout the corpus. Simply stated, Tf-idf rules out common words and highlights those that are rare. Considering the presence of very similar documents in the natural archive, and with the goal of highlighting the subtle differences between these very similar documents in distinguishing topics that occupied the different staff members, I

used Tf-idf weighting to construct the vector space model. For this, the Tf-idf formula was included in the software developed for the rest of this study. This means that the first step involved in processing the absolute frequency .txt matrices obtained from Rainbow was the Tf-idf calculation.

Cosine similarity

The pairwise distance or similarity between each vector (document) in the vector space model is calculated using the cosine similarity distance formula. This is a geometric calculation of the cosine of the angle formed by the two vectors that represent the documents whose similarity needs to be determined.[172] The formula considers the normalized or weighted frequency of each word in a vector/document in comparison with the frequency of the same terms in the other vector/document and in relation to the length of both vectors. Cosine similarity results range from 1 to 0. A result of 1 means that two documents are identical: the closer the number to 1, the more similar the documents. Table 2 shows a symmetric matrix containing the cosine similarity measures for seven documents belonging to three different authors (cultural manager, director, president). Observe that the diagonal line shows cosine similarities of 1 since it indicates the comparison of the same document to itself. In this example, documents 5 and 4 (A999123.txt and A999034.txt) are the most similar.

Table 2: Matrix of cosine similarities between 7 documents.

Documents	Cultural manager 1	Cultural manager 2	Cultural manager 3	Director 4	Director 5	President 6	President 7
Cultural manager 1	1	0.229	0.075	0.161	0.117	0.068	0.224
Cultural manager 2	0.229	1	0.205	0.406	0.292	0.23	0.077
Cultural manager 3	0.075	0.205	1	0.183	0.088	0.091	0.058
Director 4	0.161	0.0406	0.183	1	**0.505**	0.313	0.085
Director 5	0.117	0.292	0.088	**0.505**	1	0.263	0.056
President 6	0.068	0.23	0.091	0.313	0.263	1	0.048
President 7	0.224	0.077	0.058	0.085	0.056	0.048	1

The main difficulty in calculating the similarities between each document and every other document in a given corpus is the number of documents and word weights or frequencies that need to be computed. Computational processing of large matrices not only requires significant storage space but also substantial processing memory. After processing, the matrices containing the cosine similarities between each document and every other document are output as .txt matrices with and without the file names, and in comma delimited or tab delimited format.

Relationships between people based on relationships between texts [173]

To explore the nature and strength of the staff members' relationships as reflected in the texts that they wrote and gathered during their work, a program for calculating four formulae was developed. Each formula provides a different perspective or synthesis of the data included in the large cosine similarity matrices. The first formula—average formula—was developed a priori during the formulation of the appraisal method, and the last three were developed as a consequence of the process of analysis and interpretation to contrast with or clarify some results. After processing the large cosine similarity

matrices, the products of the calculations are printed as space delimited matrices in .txt file format. The different output files have different naming conventions, so it is possible to recognize which formula was used to create them. The use of any of these formulas can be chosen by typing the corresponding option in the program's command. Also during the process, the program generates a .txt file with the initials of the staff members involved in the yearly set that is being processed. In this way, the provenance of the records is always known. Below I describe the role of each formula. As a guide for the reader, because all the formulas derive from the "average" formula (1), many of the same variables described for it are used in the other formulas.

Averages

Through the average formula it is possible to determine staff member to staff member relationships based on document to document similarities. This is the basic formula used to determine the strength of the relationships between staff members based on the texts that they wrote and co-wrote. Any two authors who write N and M number of documents respectively will have N times M similarities in the final similarity matrix. To calculate the average cosine similarities between two authors, all the N times M similarities are added and divided by N times M. In this way, a symmetric average matrix can be obtained using formula (1).

$$Ave(j_j, j) = \frac{\sum_{x=F_j}^{L_j} \sum_{y=F_j}^{L_j} C(x,y) - (L_j - F_j + 1)}{(L_j - F_j + 1)(L_j - F_i)} \qquad (1)$$

In formula (1) C (x, y) is the similarity between document x and document y. Suppose that the document number for author *j* starts from F_j (first) and ends with L_j (last), the document number for author *i* starts from F_i and ends at L_i. Then the formula excludes the diagonal "1"s or self similarities for the average calculation.

Table 3 below shows the matrix with averages of cosine similarities between pairs of staff members for the year 1997. The blanks in the diagonal cells correspond to averages between documents of the same staff member, which are not considered for interpretation purposes. The matrix allows various analyses and comparisons. The averages (bolded in blue) corresponding to the director suggest that the majority of the staff members are strongly related to him. In the row corresponding to the cultural assistant 2 we can observe (bolded in red) that her relationship to the projects assistant is equal to the one with the cultural assistant 1, and both are higher than her relationship with her supervisor, the cultural manager (bolded in black). We can also observe that compared to everybody else, the projects assistant has the weakest range of relationships (bolded in green) with the rest of her co-workers.

Table 3: Averages of cosine similarities between pairs of staff members, year 1997

	No. of records	cultural manger	cultural assistant2	financial manager	projects' assistant	social manager	director	director's assistant	cultural assistant1
cultural manager	96		0.026	0.019	0.015	0.020	0.036	0.023	0.023
cultural assistant2	85	**0.026**		0.022	**0.029**	0.022	0.034	0.032	**0.029**
financial manager	21	0.019	0.022		0.015	0.023	0.025	0.021	0.017
projects assistant	25	**0.015**	0.029	**0.015**		**0.016**	**0.018**	**0.016**	0.023
social manager	89	0.020	0.022	0.023	0.016		0.026	0.017	0.016
director	80	**0.036**	**0.034**	**0.025**	0.018	**0.026**		**0.043**	0.023
director's assistant	206	0.023	0.032	0.021	0.016	0.017	0.043		0.020
cultural assistant1	124	0.023	0.029	0.017	0.023	0.016	0.023	0.020	

In the natural archive the number of records held by each staff member is uneven and their contents vary. Therefore, average results depend on how the variables "strength of similarity between records" and "quantity of records" combines: basically the influence in which each of these variables exists in each relationship. For example: in Table 3 above, results indicate that during that year the director was a central figure in the organization reflecting the fact that in his virtual folder he kept a majority of records related to each of the areas in the foundation. To better understand the combination of quantity and similarity between records and to obtain different views of the relationships, other formulas were developed.

Sum of cosine similarities [174]

The "sum of cosine similarities formula" was experimented with to better understand the relationship between quantity of records and similarity in contrast with the results provided by the average formula. For this, the cosine similarities involved in each pairwise relationship are added. The sum formula (like all the formulas included in the computer program) is based on the average formula. It is as follows:

$$R(j,i) = Ave(j,i) \cdot (L_j - F_j + 1) \cdot (L_i - F_i + 1) \qquad (2)$$

in which, R is the new sum matrix, Ave matrix is the average matrix that is calculated by formula (1), (L_j - F_j+1) is the number of documents by author *j* and (L_i - F_i+1) is the number of documents written by author *i*. Table 4 below, generated with the sum of cosine similarities formula, shows that the strongest relationships (bolded) correspond to the staff member who has more records in her folder.

Table 4: Sum of the cosine similarities between every other staff member, year 1997.

	No. of records	cultural manager	cultural assistant2	financial manager	projects' assistant	social manager	director	director's assistant	cultural assistant1
cultural manager	96		206	38	35	159	267	444	262
cultural assistant2	85	206		41	63	168	231	564	311
financial manager	21	38	41		8	44	43	95	45
projects' assistant	25	35	63	8		35	37	84	71
social manager	89	159	168	44	35		185	316	175
director	80	267	231	43	37	185		723	233
director's assistant	**206**	**444**	**564**	**95**	**84**	**316**	**723**		**536**
cultural assistant1	124	262	311	45	71	175	233	536	

In this table the strongest relationships are between the director's assistant—with more records than anybody in the set—and the rest of the staff members. When relationships are based on the sum of the cosine similarities, results are influenced by quantity of records. Still, similarity will be influential depending on the proportion of similar records included in a given relationship. This is clear in the result obtained in the relationship between the director's assistant and the director which is also reflected in the average results in Table 3. The results shown in this table are consistent with those obtained across the yearly sets.

Balanced averages

An intermediate option between the average and the summation of cosine similarities is the "balanced average." The goal of this formula is to lower the influence of the variable quantity of records involved in any given relationship based on the average calculations. Results are obtained by dividing the summation of cosine similarities by the multiplication of the square roots of the numbers of documents written by each staff member involved. The formula is as follows:

$$R(j,i) = Ave(j,i) \cdot \sqrt{(L_j - F_j + 1) \cdot (L_i - F_i + 1)} \qquad (3)$$

What is referred as *Ave* is the average matrix calculated by formula (1). (L_j - F_j+1) is the number of documents by staff member *j* and (L_i - F_i+1) is the number of documents written by staff member *i*. The use of the square roots lowers the influence of the total number of records involved in a given relationship as the square root of a number is lower than the number itself. Table 5 below shows the results obtained using the balanced averages for 1997. Compared with Table 4 we observe that even though results are less impacted by the total number of records involved, the influence is still significant. In this case, because the director's assistant has the largest number of records she has a majority of the strongest relationships in the matrix.

Table 5: Balanced average of cosine similarities between pairwise staff members, year 1997.

	No. of records	cultural manager	cultural assistant2	financial manager	projects' assistant	social manager	director	director's assistant	cultural assistant1
cultural manager	96		2.34	0.86	0.73	1.78	3.13	3.24	2.47
cultural assistant2	85	2.34		0.96	**1.37**	1.95	2.80	4.26	3.03
financial manager	21	0.86	0.96		0.36	1.01	1.04	1.44	0.87
projects' assistant	25	0.73	1.37	0.36		0.74	0.82	1.17	1.28
social manager	89	1.78	1.95	1.01	0.74		2.21	2.36	1.68
Director	80	**3.13**	2.80	1.04	0.82	2.21		**5.63**	2.34
director's assistant	**206**	2.24	**4.26**	**1.44**	1.17	**2.36**	**5.63**		**3.35**
cultural assistant1	124	2.47	3.03	0.87	1.28	1.68	2.34	3.35	

After analyzing the results in the different sets I decided not to use the sum or the balanced average formulas and concluded that the average formula better resolves, across the sets, the varied combinations between similarity and quantity of records. And yet, these exploratory experiments allowed a better comprehension of the characteristics of the corpus and the problems faced when dealing with irregular sets like the ones in this archive.

Relative averages

As I was analyzing the average matrices I observed that different staff members had different ranges of results (this can be observed by looking at the range of results on the individual rows). Some staff members had consistently overall high relationships with

the rest and others overall medium and/or low relationships. This drew my attention to viewing results from the perspective of each staff member. The average matrices are symmetric, which means that the relationship between A and B is equal to the relationship between B and A. Average matrices give an organization-wide view of relationships in which results are compared in relation to everybody else. Instead I wanted to find out what the relationship AB would mean to A and what to B. When looking at relative averages you focus on the results of the row that corresponds to one staff member and his or her relationships with the rest.

Relative averages are obtained by multiplying each of the averages of cosine similarities from a given row by a different factor. These factors are calculated by dividing 1 by the average of each row (excluding the numbers in the diagonal which are self similarities). Using this formula the averages of all the matrices are 1. The formula is as follows:

$$Rowave(j) = [\sum_{i=1}^{N} Ave(j,i) - Ave(j,j)]/(A-1) \qquad (5)$$

in which A is the number of authors, and the Ave matrix is the average matrix that is calculated by formula (1), and *j* and *i* are the different staff members. As a result, the matrix obtained is not symmetric.

Table 6 below shows the different way in which averages (first two) and relative averages (bottom two) render relationships for two staff members for the year 1997. To appreciate the differences between these two views of the data, I bolded averages higher than 0.024—average of average matrix—in the average rows and averages higher than 1—average of relative average matrix—in the relative averages rows. I use 0.024 and 1 as thresholds to signal that above average relationships are stronger. In this way it can be observed that the relative averages render more above average relationships for each staff

member, basically because the threshold results from the individual staff member's range of relationships. It can also be observed that when viewed from the relative perspective, relationships have a different result for each staff member involved in the relationship. In the example below, the relationship to her supervisor (the cultural manager) is stronger for the cultural assistant 1 than the reverse.

Table 6: Comparison of results from averages and relative averages for two staff members, year 1997.

Averages

	No. of records	cultural manager	cultural assistant2	financial manager	projects' assistant	social manager	director	director's assistant	cultural assistant1
cultural manager	96		**0.026**	0.019	0.015	0.020	**0.036**	0.023	0.023
cultural assistant 1	124	0.023	**0.029**	0.017	0.023	0.016	0.023	0.020	

Relative averages

	No. of records	cultural manager	cultural assistant2	financial manager	projects assistant	social manager	director	director's assistant	cultural assistant1
cultural manager	96		**1.12**	0.83	0.65	0.84	**1.15**	**1.00**	0.98
cultural assistant 1	124	**1.05**	**1.34**	0.78	**1.04**	0.74	**1.07**	0.95	

In the context of a natural archive in which the quantities and types of records stored by staff members are so disparate and change from one year to the other, the relative average results are complementary to the average ones. The latter provide a rendition of organization-wide relationships while relative averages express relationships

maintained by each staff member with the rest of the staff members within his or her range of average cosine similarities. The results above are consistent in both views, the highest relationships—cultural manager with director and cultural assistant 1 with cultural assistant 2—agree.

REPRESENTATIONS

As has been explained, the vector space model for each yearly set is a numeric representation of the records included in the set. In this study I had to consider whether the records were accurately represented, which relates to archival concepts of integrity and authenticity. One way to look into this problem was to compare yearly matrices obtained from a non-stemmed vector space model with those obtained from a stemmed vector space model. Another possibility was to generate average matrices from which low cosine similarities were removed from the calculation.[175] In both cases, the idea was to stress the similarities between records by narrowing the vector space model in the first case and by narrowing the differences between staff members in the second case.

To stem or not to stem

The behavior of the stemmer was evaluated by comparing stemmed and non stemmed average matrices to determine if and how this process affects the relationships between staff members. This implied analyzing the numeric matrices, the network diagrams before and after stemming (network diagrams with and without stemming can be observed in Appendix III), and determining whether the results agree with the accounts of the staff members. Particularly in this corpus in which so many words and their variations are repeated across the archive, I was concerned that stemming could flatten the representation and dim the differences between texts. Also, there are many

texts that have identical parts due to cut and pasted fragments and the use of documents as templates (part of the shared records creating practices), and therefore the risk is that by stemming, the identity of those staff members who kept those records can be obscured, which is the opposite of what I want to observe through the relationships.

For example we can consider the ubiquitous uses of the words common to all areas, such as project, foundation, grant, beneficiaries, meeting, subsidy, and their respective variations; when stemmed, they will engender similarities between records from different members that are not based on projects that they worked on. Table 7 shows a comparison between stemmed and non stemmed results for two staff members during the year 1999. To convey the impact of the stemmer I rated the relationships in each section from strongest to weakest (1 to 6).

Table 7: Comparison of non-stemmed and stemmed sets, year 1999.

NON-STEMMED	cultural manager	cultural assistant2	**STEMMED**	cultural manager	cultural assistant2
cultural manager		**(3)0.0259047**	cultural manager		**(2) 0.048238**
cultural assistant2	**(6) 0.0259047**		cultural assistant2	**(5)0.048238**	0.069969
social manager	**(4) 0.0265573**	0.021279	social manager	**(2)0.05172**	**(6)0.039398**
financial manager	0.019289	0.018456	financial manager	0.032959	0.028185
projects' assistant	0.007887	0.011875	projects' assistant	0.012566	0.016565
president's assistant	**(2) 0.0290725**	0.012791	president's assistant	**(6)0.046908**	0.019658
receptionist1	0.010180	0.010344	receptionist1	0.019208	0.017958
science assistant	0.017159	**(2)0.0260684**	science assistant	0.033231	**(3) 0.042058**
president	0.025346	**(5)0.022146**	president	**(4)0.048691**	**(5)0.04017**
director	**(1) 0.0489564**	**(1)0.0430502**	director	**(1)0.087094**	**(1)0.070447**
director's assistant	**(3) 0.0280854**	**(4)0.0245931**	director's assistant	**(3)0.050437**	**(4)0.040986**
cultural assistant1	**(5)0.0262492**	**(6)0.0219072**	cultural assistant1	0.045905	0.035375
cultural assistant3	0.022698	0.011462	cultural assistant3	0.039538	0.019234

The results show that while some relationships maintain their order in regard to the non-stemmed set, others change it slightly, and others drastically. The relationship between the cultural assistant 2 and the social manager moves from number four in the non-stemmed set to number two in the stemmed set. Also, the relationship between the cultural manager and the president's assistant changes from number two in the non-stemmed set to number six in the stemmed one.

I decided against using the stemmer because I concluded that in this corpus in which so many words are shared across documents, depending on the nature of the records involved in the relationship, the stemmer could emphasize similarities based on the use of high frequency words and families of words but not necessarily on similar

projects or collaboration. Therefore, I considered that it was more accurate at this point and in this context to base the method on the basic, non-stemmed yearly sets. Furthermore, through these exercises I understood that while this appraisal method borrows methods from IR, it is not an IR problem permitting dimension reduction. Being an archival problem, the integrity of the records representation has to be considered. The stemmer behaves very obscurely in this archive full of repetitions in which its consequences are uncertain. Besides, the question about computational resources needed to process larger matrices can be answered through allocations in super-computer centers without having to resort to stemmers simply to secure dimension reduction.

Filtering low cosine similarities

With the goal of eliminating relationships between records that have little or no similarity, I explored filtering techniques. For this, code was written into the program to ignore specified ranges of cosine similarities between records in the calculation of the pair-wise averages between staff members. So, for example, I could choose to discard cosine similarities lower than 0.01 (which is almost no similarity at all) from the calculations. In this filtering method, records are not removed from the model, just the weak cosine similarities or relationships. So for example if a given relationship between record A and record B is weak and is not counted towards an average, the relationship between A and C does count because it is above 0.001. Partial views (a few staff members to all staff members) of the matrices corresponding to the year 2001 in Table 8 below show the different results obtained with and without filtering. To better illustrate how tight a filtered relationship can get, I used the average of the matrices as thresholds to show above and below average similarities. The non-filtered matrix has an average of 0.026, and the filtered one of 0.024. Marked in red in the non-filtered matrix are

relationships above average. The bolded red results in the filtered matrix show relationships that are not above average in the non-filtered matrix but become above average after filtering.

Table 8: Comparisons of results between cosine similarities and filtered cosine similarities in the calculation of averages, year2001.

Non-filtered cosine similarity matrix with average of 0.026

	social manager	receptionist2	president's assistant	receptionist1	president	director	director's assistant	cultural assistant1
projects' assistant	0.017488	0.020371	0.028137	0.025391	0.036976	0.026129	0.023992	0.022444
science assistant	0.016240	0.035712	0.017467	0.027969	0.037927	0.025921	0.024930	0.014536
social assistant	0.055115	0.018028	0.016312	0.020921	0.035746	0.023673	0.023690	0.017112

Filtered cosine similarity matrix with average of 0.024

	social manager	receptionist2	president's assistant	receptionist1	president	director	director's assistant	cultural assistant1
projects' assistant	0.019260	0.021668	0.031162	**0.027257**	0.038996	0.027367	**0.030203**	0.025948
science assistant	0.017452	0.037078	0.019290	0.029845	0.039598	**0.026745**	**0.030068**	0.016504
social assistant	0.058697	0.019396	0.018615	0.023273	0.038086	**0.025168**	**0.031256**	0.019555

Using the cosine similarities filtering option, all the relationships between staff members tighten and some relationships that in the non-filtered option are below or close to average become above average. I decided against using the filtering option because by discarding from the average calculation low similarities between records, I could be artificially narrowing the relationships between staff members who have many records with nothing in common. That is, I was stressing similarities but not dissimilarities. The results have the same archival drawbacks of being an altered representation of the documents.

In the future, a better way to approach filtering would be to calculate the percentage of times in which a given record shows low similarity with the majority of the records in the set. If the percentage is very high, this could mean that the record is not related to the actions and activities in the organization, which in archival terms is called a non-record and in IR is called noise. To be precise this technique would have to be complemented with corroboration of the contents of the texts involved and include testing with different levels of cosine similarities. In turn, the exercise might help establishing what a non-record is in a natural archive, whether it is a short record or a record that contains unrelated content.

Formulas

After testing, comparing, and documenting the different alternatives and their results, including stemmed and non stemmed sets; sets treated with filtering and stemming options and with and without filtering; and using averages, summation, and balanced and relative averages—for the final appraisal method I used the following combination of formulas:

- Vector space model: cosine similarities for all pairs of records, with Tf-idf and a non stemmed bag of words representation
- Relationships between staff members: averages of cosine similarities between records of pairs of employees and relative averages of cosine similarities between records of pairs of employees for verification purposes

MODIFYING AND BUILDING A TEXT MINING PROGRAM

One of the most rewarding processes of this research was building the appraisal text mining program. While not a common activity for archivists now, creating software to analyze and organize archives will become so in the near future, so I want to convey what this experience involved and meant to me. As I have mentioned before, Dr. Hai Bi was responsible for modifying Rainbow to build an appropriate bag of words representation and for coding the rest of the program. My participation included both administrative and software development tasks. For the former and through my dissertation chair, I requested a small research grant from the Office of the Vice President for Research at University of Texas at Austin. For the latter, I researched formulas and open source tools, tested the program, reported bugs, analyzed the results, and discussed with Hai Bi options to improve or modify the program.

As I provided the Tf-idf and cosine similarity formulas found in the bibliography, Hai Bi customized them to combine my needs and the characteristics of the data at hand. For example, to calculate averages of cosine similarities between pairs of staff members, his algorithm had to point to the specific identifiers that indicated which records belonged to which staff member. For this he used the staff member's initials present in the raw frequency matrixes output by Rainbow. In turn, this initial belongs to the virtual folder in which the records belonging to that staff member are stored. So, if the file path is **/1991/A1/A1990002.txt/A1,** the indicator of the file's provenance is A1.[176]

Testing and debugging the program allowed us to figure out adjustments that needed to be made. For example, we had to deal with the processing power limitations of the server that I was using. To deal with these problems, Dr. Hai Bi wrote code that distributes computing resources by allowing data to be stored in the hard-disk as

calculations take place. Because this causes delays by increasing the processing time, he changed the order of the formulas according to my needs so that I could obtain certain outputs first and stop the process at any time afterwards depending on results. He wrote this code in such a way that if the program is used on a more powerful server, we can choose a command that allows processing to go faster. Another flexibility feature built into the program is that it offers the option to output matrices as comma or space delimited. Also, cosine similarity matrices can be generated with and without the names of the text files involved in the calculation. These options facilitate using the data with different visualization software or other software to further analyze or process the results.

Formulas specifically developed for this method, such as the cosine similarity filtering option, the different ways of calculating averages, or the option to calculate the sums of cosine similarities were a consequence of extensive testing and analysis of results. Some functions turned out to be very useful, while others were not used, but in the process, we were learning. Modifying open source software and creating new text mining software was and still is a learning process, and I was very fortunate to find an enthusiastic, patient, and knowledgeable partner with whom I shared language and computing barriers. Hai Bi is Chinese, and he had to learn about Spanish accents, diacritics and inflections to modify Rainbow's tokenizer and to add and test the stemmer. We met on various occasions so that he could explain to me the mathematics included in the code. We worked together through language barriers, computing limitations, and my math deficiency.

The program uses the frequency matrices obtained from Rainbow to calculate Tf-idf weighted frequencies for each document and then outputs cosine similarities between all pairs of documents in a given matrix. Finally, it produces averages—normal,

balanced, and relative—and sums of cosine similarities between all pairs of staff members. It also has filtering options to discard ranges of cosine similarities. For each of these functions the program outputs different .txt files that can be identified through their different naming convention. It is now on its twentieth version.

VISUALIZATION AND INTERPRETATION

During one of the sessions of the Digital Humanities conference that I attended in June of 2007 I heard a presenter say more or less, "Visualization is not a consequence of research; it is the research."[177] I find that this concept is applicable to this work because as the research progresses and results are visualized, doubts and new ideas emerge. It is difficult to interpret results from numeric matrices, as one can only focus on a few results at a time, making it difficult to identify changes or patterns. To interpret the text mining results, I use social network analysis diagrams and animations. Both methods provide complementary views. The interpretations in this dissertation are derived from the social network analysis diagrams, since the animations and interactive visualizations are still in development.

Social network analysis

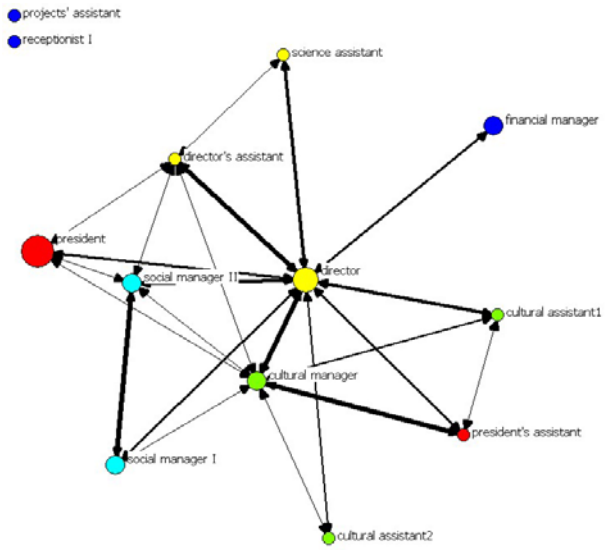
Social network analysis has been applied to organizations for thirty years.[178] It focuses on studying relationships between actors holistically, emphasizing the observation of the structure of the network in which the actors are included and interact.[179] Rather than with the strength of relations between actors, often these studies are concerned with the connectivity of the network based on survey results about who talks or works with whom.[180] In this research I propose to perform such analysis by deriving network topology from the digital texts of the organization. To visualize and interpret the

text mining results of each yearly set I used UCINET; a social network analysis software that produces static diagrams of the average matrices.[181]

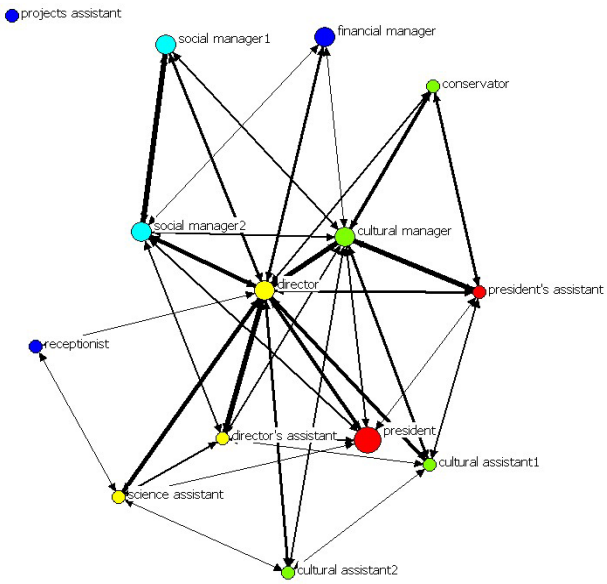
One of the main assumptions in social network analysis is that the position of an actor in the network is determined by his or her relationships with everybody else. So for example, when an actor leaves the network, or his relationships with others change, the entire configuration of the network changes. To illustrate this phenomenon and demonstrate how UCINET's diagrams represent the data, Figure 4 below shows three graphics corresponding to the year 1998. The connecting lines are called *ties* in social network analysis jargon, and their thickness shows the degree of *strength* in the relationships. The *nodes* are in this case the staff members; nodes with similar colors denote common functional areas, and their size the hierarchical status in the organization. The color coding is as follows: yellow for the director's office and the science and education areas, green for the cultural area, blue for the financial area, red for the president's area and turquoise for the social welfare area. In terms of size, the higher the hierarchical position the bigger the size of the node.

The position of each node in the network depends on the number of relationships/ties to other nodes and their strength, denoted in the drawing by the thickness of the tie. So people with many relationships and stronger relationships are drawn to the middle of the network, and people with fewer ties, or with ties to the outside of the group being studied, will be located closer to or in the outskirts (the analogy that can be used here is that a person is being yanked to its position in the network by the strength and number of relationships). This in turn means that the numeric average of cosine similarities is not represented as a distance in the graphic. Diagram (a) has 13 members and diagram (b) has 14. When one staff member is added or removed from a

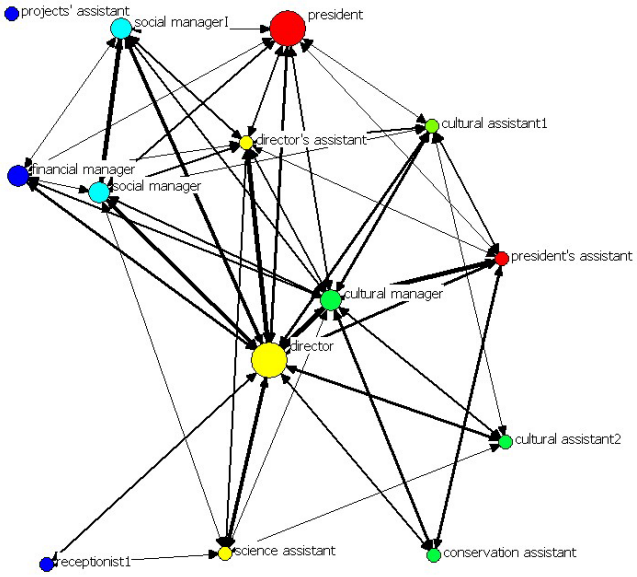
given set the vocabulary in the bag of words representation increases or decreases respectively and the results of the average cosine similarities change. Diagram (c) shows a matrix of 14 members whose vocabulary was stemmed and as a consequence, the number of relationships between staff members increased and their positions in the network changed.



a) Network drawing without conservation assistant



b) Network drawing with conservation assistant



c) Network drawing with conservation assistant and after stemming.

Figure 4: Network diagrams of the averages of cosine similarities generated with UCINET, year1998.

Thresholds

The study of social network structures can be based on binary relations such as whether actors are or are not related to one or to many, or who reports to whom. They can also be based on other types of quantifiable and continuous data that expresses degrees of relationships. In this research, relationships are based on the similarities between the texts that people wrote and gathered with the average results between pairs of staff members expressing the degree to which those relationships based on documents existed at a given time.

UCINET gives the option not to show ties whose values fall below a certain number so that the generated diagrams will render relationships above a specified threshold. This has a practical side, as the diagrams are not clear when everybody is related to everyone else, which in a matrix based on use of common words is a frequent outcome. So, the challenge resides in choosing this threshold and finding out what it means. Another related complexity is that in organizations there are not only weak and strong relationships between staff members but also different levels of relationships. For example, the relationships between managers with similar functions, between managers and their assistants, and between managers and the assistants of other managers will not be the same. In turn, all of the above varies depending on whether the records generated are more or less structured and controlled.

I tested several ways of choosing a threshold. One of the first methods was to input the average matrix data in a spreadsheet, generate a curve with all the averages, and choose the point in which the curve denotes a significant change. Figure 5 below shows the chart of the year 1996. In it, the curve shows a gradual ascent so it would be arbitrary to choose any point between averages of 0.015 and 0.035. On the other hand, if I select

the point 0.035, where there is an obvious leap, I would be highlighting only the strongest and very few relationships. I concluded that this method was ambiguous, and that the decision to choose a certain point for each yearly set was not going to be consistent throughout the different samples.

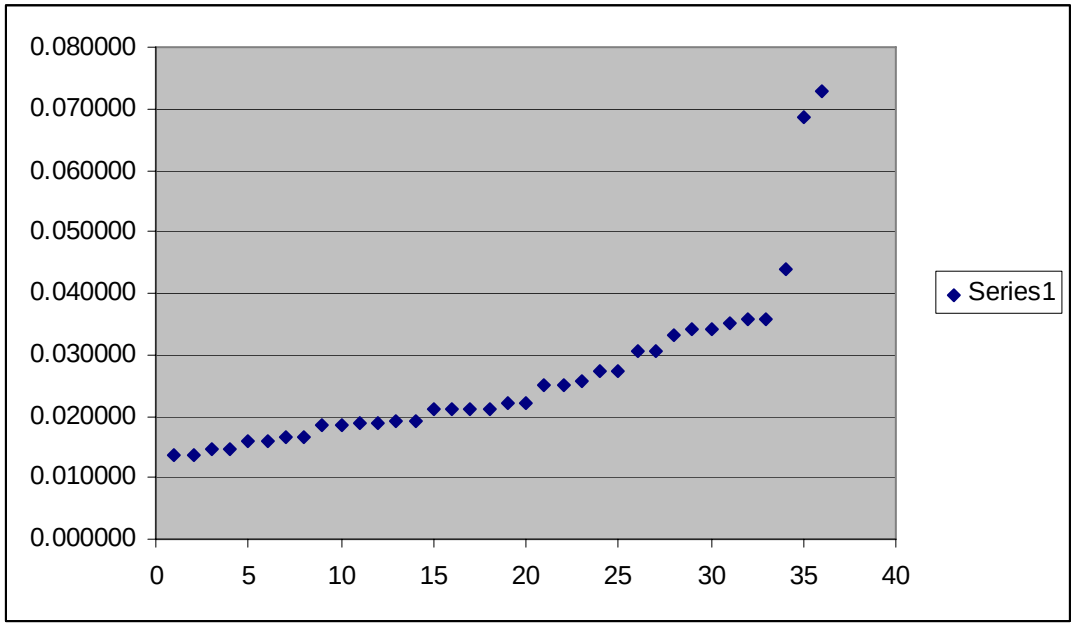


Figure 5: Averages of cosine similarities, year 1996.

An efficient and somewhat problematic way to calculate the threshold is by averaging the results in the matrix, excluding the self similarities present in the diagonal. This calculation can be applied consistently throughout the sets. The drawback with this method is that depending on the results obtained in the different yearly matrices, averages are biased by high numbers or strong relationships and therefore exclude low level relationships. To compensate for this problem, I used the relative average calculation to understand what the relationship with his or her co-worker means to an individual staff members, which will be further explained in the results sections. The diagrams in Figure

6 below show the differences between visualizing the matrix without a threshold (1) and with the threshold (2). In this case in which everybody shares the same language and is related to some degree to the rest, it is not useful to observe the matrices without thresholds. As the numeric threshold increases, only the strongest relationships are rendered. If the strength agrees with the results of the interviews, the rendered relationship is meaningful.

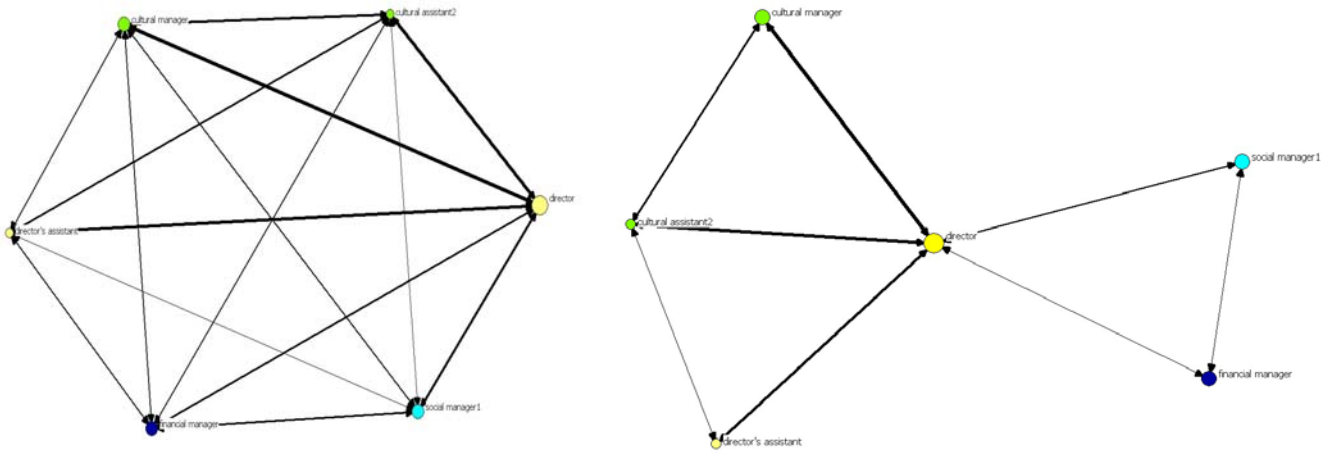


Figure 6: Incidence of the use of a threshold, year 1996.

The issue of defining a threshold is not just a feature of UCINET software. Finding a threshold that defines the nature of the relationships between staff members (based on projects and themes, or based on use of common words) was one of the expectations of this study. However, finding a consistent threshold in a natural archive in which each yearly set is different from every other (and even each set of relationships is different from the next) was not possible. Thresholds are unique to each yearly set and cannot be transferred to the other sets.

A problem with the thresholds that I am using in this method is that they draw all the relationships to a mean. But just as in real life, relationships are multi-layered and

different people can have different thresholds depending on how many records and the types of records that they wrote because of their function in the organization. An alternative view is to focus on individual staff members and their respective one-to-many relationships in which the average cosine similarities are visualized as real distances in a space.

Animation

With the static diagrams I could only compare one year—or one static diagram—to one year at a time and had difficulties obtaining a comprehensive picture of the changes that occurred during the ten years. Work-dynamics through time can be better visualized and analyzed through animation. To obtain help making an animation, I contacted the Visualization Team from the Texas Advanced Computing Center (TACC), a research group at the University of Texas at Austin that consults with researchers to create scientific visualizations.

The collaboration with the visualization scientists started by discussing the research goals, the characteristics of the data at hand and the options available to represent it as well. The first challenge was imagining how to show the data, considering that the metaphors used to represent it should aid interpretation. I thought that the process must be similar to the one that animation artists go through to create a movie from a story and immediately realized how hard that is. We decided to start with a simple project and visualize the relationships between one staff member and the rest of the staff members through time, and to consider the full range of relations including the weak ones.[182]

I chose to animate the director's relationships because of his role in the organization's records-creation process. The data needed to create the animation comprises the row containing the averages of cosine similarities that corresponds to the

relationships between the director and the staff members from each of the 10 yearly matrices. The data sets were provided to the visualization scientists labeled with the staff members' roles and with an explanation of the functional areas in which each of them worked. It had to be provided in a format that can be converted to the native visualization software format. For this project I formatted it as an Excel spreadsheet. The director's visualization was done in ParaView, an open source visualization program.[183]

Figure 7 below shows screenshots of two moments in the animation. Continuing with the style of the static diagrams, staff members are represented as spheres, colors indicate functional areas, and sizes indicate hierarchies. The director is at the center, and the co-workers are spaced around at distances scaled by the inverse of the cosine-similarity values for that year which means that the distances reflect the strength of the relationships in comparison to each other. In the animation, when someone leaves the organization, they fly off-screen and the opposite occurs when someone enters. The frames between the years are linear interpolations which can be increased or decreased to achieve a smoother or more rapid animation. As transitions between years occur, the spheres get labels of who is who and the animation stops briefly (or it can be stopped) to allow analysis.

Using this visualization I was able to observe who maintained steady relationships with the director, who got closer to him over the years, who from each area worked closely with him and when, who worked with him until he left the organization, and how his role changed during his last year in the organization. The animation, which can be viewed with Real Player, is included as a supplemental file called Use of Text Mining and Visualization to Infer Work Dynamics from Organizational Records.

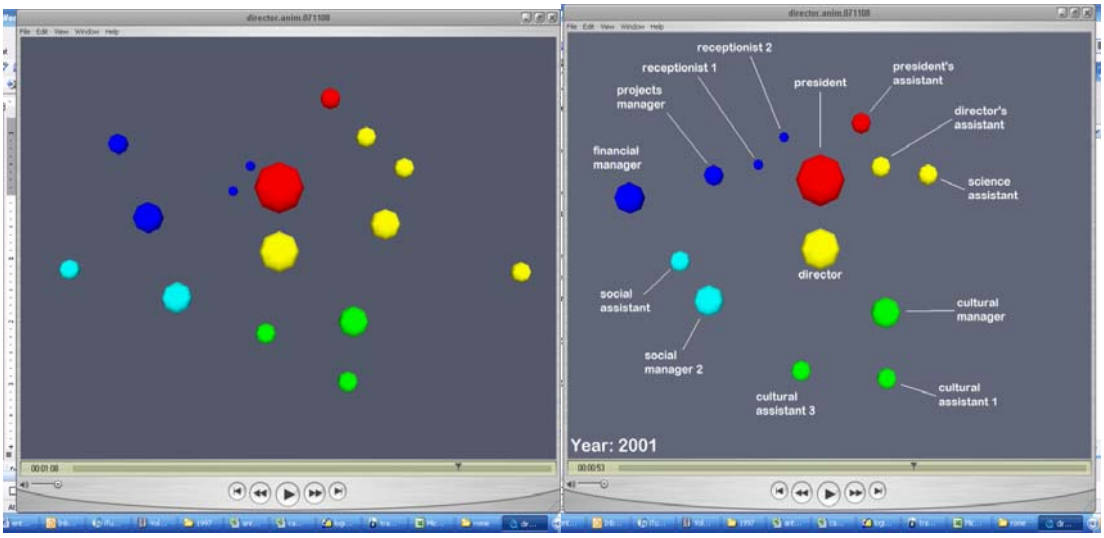


Figure 7: Screenshots of animated visualization of the relationships between the director and the rest of the staff members, years 1996 to 2004

Overall, the possibility of analyzing the information beyond the individual details provided by the yearly network diagrams shows a broader view of the organization as people entered, worked, and left. As observed by Dr. Victoria Horwitz, one of my dissertation committee members, the animation shows the lack of stability, the constant reconfiguration of the roles in the organization, in contrast to organizational chart models that separate the strict incumbencies between staff members.[184]

Myself

Having worked in the organization in the past, throughout the interpretation process I had to distance myself from my perception about the institution, my tendency to direct my attention towards the area that I worked on, and my memories about the relationships among people. Aleph was a relatively small organization, and while most people knew what was happening in general, not everybody knew in depth what their co-workers in other areas and sometimes in their own area were doing. Most importantly, I

had to remember that while people could know what their co-worker was doing, they did not necessarily share the same records nor work on the same project.

VALIDATION

External validity

Validating the results obtained through the appraisal method means that they should reflect the organization's work dynamics. In this study, work dynamics are understood as the way in which staff members at Aleph entered and left the organization, whether they fulfilled one or dual roles, changed them temporarily or permanently, how people from different functional areas grouped to work on specific projects, and how they worked with each other and with the community that they served. In this context, work dynamics can be strict, loose, and ever-changing; involve hierarchies that are or are not respected or enacted; and reflect people helping each other or replacing someone while on sick leave.

To confirm results I used various methods. I conducted interviews with 13 staff members—65%—whose records were stored on the shared drive and with the systems administrator who managed the networked server (Appendix I includes the questionnaire protocol).[185] During the interviews I specifically asked about roles and functions to capture the variety of activities that staff members completed during their tenure and whether it changed and at what point. I also asked them about work processes and individual record-keeping practices which, included in the narrative of the archive's formation process, explain gaps in the archive, version control, and retention or disposal practices. Therefore, based on the results of the text mining process and in contrast with

the data gathered in the organization, I make inferences that are supported by these different sources, just as I did in the formation process study.

But the method is also validated when results reflect known truths, for example, that the appearance in the yearly graphics of new staff members and the exit of those leaving agree with the dates of their tenure are valid results. That the welfare manager is consistently tied to her assistant is also a positive result; and that the financial manager's ties are inconsistent through the years reflects the fact that he rarely generated or gathered texts: he mostly used spreadsheets and the financial systems to conduct his work.

About the results that could not be validated, either because there are gaps in the archive that I do not know about or because there are events that the interviewees did not remember or I failed to ask about, I return to the idea of the archaeological site. I make inferences about what is there and what I can observe, and it is valid to do so as long as the limitations for explaining those gaps are exposed. When in doubt about the reasons behind the strength or the weakness of some relationships, I reviewed the contents of the records involved in the set. For example, the presence or absence in the staff members' folders of similar record versions or records that treat similar topics or similar record types explains and validates results. However, I have to clarify that these observations were not exhaustive, as I could only verify a random number of records and not every single one. In other cases I conducted a cluster analysis of the records (that is, I grouped documents at similar level of cosine similarities) to understand which records were driving the similarities or the distances.

I also re-interviewed some staff members by email and face to face. With the purpose of clearing some doubts, in the winter of 2006 I brought the network diagrams to Buenos Aires and showed them to a group of staff members. Looking at the diagrams

they were able to provide answers to most of my questions. However, we focused our observations on certain problems and did not analyze the graphics as a whole. In sum, some results could not be validated through interviewing, but I pursued ways to explain them through the methods mentioned above.

Internal validity[186]

In this appraisal method each and every relationship is different from the rest. One approach to test the internal validity of the average results—considered expressions of the strength of the work relationship between co-workers—is to produce distribution curves of cosine similarities involved in given relationships and see whether they match the average results as a comparison of one year to the next. The internal validity curves that I present in this research are an ongoing exploration. So far, the results obtained through the distribution curves agree with the average results, but I have not yet exhausted the tests for each and every relationship and every year. The distribution curves also provide other ways of analyzing and visualizing the data produced.

The way in which internal validation was approached was by identifying average results for which I did not have an external validity explanation, or that showed abrupt differences with earlier patterns: for example, a sudden distance between two staff members who had consistently worked together in past years. In this part of the project I also worked with Dr. Hai Bi, who built into the text mining program the possibility of obtaining sub-matrices of cosine similarities between staff members for any given yearly set and helped me find the adequate formula for the distribution curves. Two distribution curves are presented in the section *Individual Relationships*.

Gold or Coal?

PROSPECTING

In this section, I present different and representative results and their interpretation. For this I use the network diagrams and the corresponding matrices. To disambiguate and better interpret certain results I use complementary tools such as relative averages, cluster analysis, and filtering of staff members from a given set. The complete set of ten yearly diagrams from the years 1996 to 2005 is included in Appendix III.

IN THE VEIN

The diagram in Figure 8 corresponds to the set for the year 1998, which includes 1199 records.

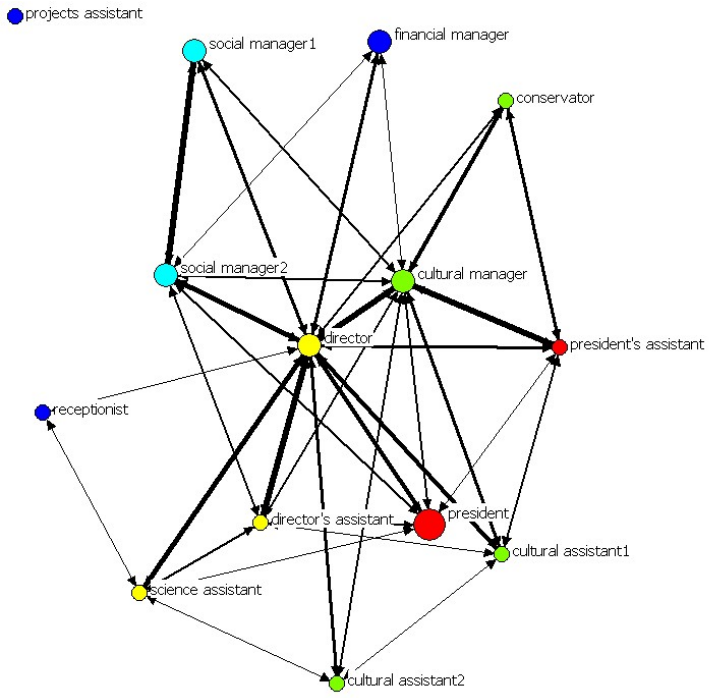


Figure 8: Network diagram, year 1998.

At the center of the network is the director, who at the time was also manager of the science and education area and edited all the official documentation generated in the organization. In accordance with his executive role, he has strong ties with the rest of the area managers and the president. The grant programs in his area were steady and run in an orderly way by the members of his team. His closeness to the cultural manager reflects his involvement in the cultural area, the least uniform in terms of the variety and number of programs released every year. His connection with the cultural area assistants is consistent with their exchange of texts that were then merged into official documents. During the busy periods in which applications were received, and to follow up specific projects, the cultural assistant 2 and the receptionist assisted the science area. Also that

year the cultural assistant 2 started to work closely with the director in the production of a conference proceedings book.

Even though he had been in office since 1996, it was during 1998 that the president started to go to the foundation on a daily basis after retiring from his full time job elsewhere. This is the first year that his virtual folder, with very few records, appears on the shared directory. His assistant's records also appear on the shared drive for the first time that year showing strong connections with the cultural manager and area staff members.[187] Between 1997 and 1999, she worked in the production of a heritage conservation project. During my interview with her I learned that she discarded regularly general types of electronic records and emails, and kept only those that she considered relevant. In this case, the majority of the records in her virtual folder are related to her collaboration with the cultural area.

That year, at the end of July, the social welfare manager 1 left the foundation. His peripheral position in the network reflects his habit of writing work procedures and criteria specific to his area that he did not share with other staff members in the organization. His strong connection with the new area manager shows the handing over of his functions: he made a point of leaving his electronic records to her. In turn the position of the new social welfare manager 2 implies her insertion into the organization.

Results under the microscope

The threshold used to render the network diagram of 1998 is 0.020. Relationships slightly above the threshold pose the question of what they mean. This question emerged on account of the above average connection between the social welfare managers 1 and 2 and the cultural manager as these areas did not have projects in common and did not share staff members (See Figure 8 above).

To analyze the results Table 9 presents the average of cosine similarities between the cultural manager and the rest of the staff members during that year. In bold are his relationships with the social area managers and in red the rest of his above average relationships. It can be observed that those with both social welfare managers are amongst the weakest.

Table 9: Above average relationships of the cultural manager, year1998.

	cultural manager
cultural manager	
cultural assistant 2	0.0241
social manager2	**0.0252**
financial manager	0.0200
projects assistant	0.0135
president's assistant	0.0424
social manager1	**0.0235**
receptionist	0.0119
science assistant	0.0185
president	0.0239
director	0.0448
conservator	0.0334
director's assistant	0.0245
cultural assistant 1	0.0276

A review of samples of records of the three area managers confirms that the areas did not have projects in common but the manager's folders do contain common words and document types such as letters and reports which conformed to a similar style. It is the case that both areas shared projects with a third party philanthropic organization and worked on the same general areas such as training and libraries, except that the approaches and goals were different. These results need to be further explored to confirm whether similarity is due to the use of words that were commonly used by all the areas

(foundation, Buenos, Aires, projects, project, etc.) which are called “high frequency words” and if by removing them the strength of the relationship will decrease.

But not all the low level relationships have the same characteristic. The average of 0.024 between the cultural assistant 2 and the cultural manager is due to the similarity and the diversity of their records. Because of her dual role assisting the cultural and the science and education areas, the records of the cultural assistant 2 correspond to her cooperation with both. In this context, her relationship with the cultural manager loses strength, and the low average reflects that her work was shared between two areas. If we look at the range of numbers in which the maximum and minimum relationship between the cultural assistant 2 and the rest of the staff members fall in Table 10, we observe that she maintained low level relationships with all the staff members. In that range (0.13 – 0.29), her relationship with the cultural manager is the second highest and the most important one is with the director, with whom that year she worked on the education and science front and in the edition of a book for a cultural project. Her next two strong relationships correspond to her fellow assistants in the cultural and science areas. Summarizing, the low range of her relationships is due to a combination of factors: she kept a high number of records; the records’ contents were diverse; she kept many records sent to her by people outside the organization; she also kept work documents that pertained to her role as assistant and were not shared with anybody else. Still, the relationships are coherent with her dual role.

Table 10: Above average relationships of the cultural assistant 2, year 1998.

	cultural assistant2
cultural manager	**0.0241**
cultural assistant2	
social manager2	0.0167
financial manager	0.0133
projects assistant	0.0150
president's assistant	0.0158
social manager1	0.0146
receptionist	0.0105
science assistant	**0.0207**
president	0.0166
director	**0.0293**
conservator	0.0127
director's assistant	0.0185
cultural assistant1	**0.0200**

Transitions or snapshots?

The transition from the year 1998 to 1999 is smooth and the overall results are consistent with the structure of the organization and with the staff members' roles. The appearance of new hires and the absence of staff members who left the organization agree with the dates encompassed by their tenure. Figure 9 below shows the network diagram of 1999. As he gathers more records related to the different areas in the organization, the president acquires stronger ties with the staff members and specifically with the director, but he does not show an above average relation with his assistant. During 1999 the president's assistant was still working in the cultural heritage project and kept records related to that. To cooperate with her and with the cultural manager, a part-time cultural assistant 3 was hired. She would remain working in specific projects in the foundation for the next six years.

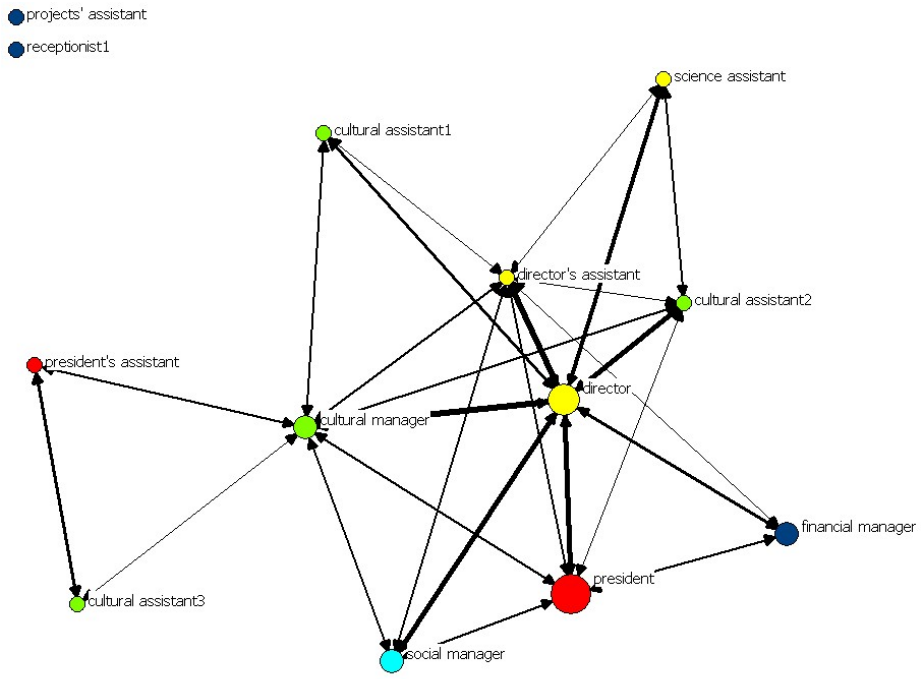


Figure 9: Network diagram, year1999.

When looking at what has changed between years it has to be noted that each yearly set varies in the number and contents of records and in the number of staff members involved. While this is exactly what happens in organizations, it has to be considered that with this visualization method each set has its own threshold driven by averages. Therefore, some relationships that were above average one year might be below average the next or vice versa.

As in 1998, all the staff members from the science and education area have above average connections with each other. This is not the same in the cultural area, in which the cultural 1 and the cultural 2 assistants are not directly connected as they were, although weakly, the year before. This result can be interpreted in light of how the work in the area was distributed, since each project assistant was assigned to different projects.

It is also consistent with what we observed before about “weak” ties reflecting their diverse occupations and their communication with people outside the organization. However, it cannot be interpreted as if they had no connection at all.

Relative averages can be used to observe what happens to relationships during transitions. Because the relative averages matrices are normalized to result in averages of 1, the transition from one year to the next can be directly compared. The relative average results cannot be graphed because UCINET only renders networked diagrams of symmetric matrices, but can be looked at in the numeric matrices. Below in Table 11 are the relative averages for the cultural assistants 1 and 2 during 1998 and 1999. Marked in bold are the relative averages for both of them and in red those that are stronger for each of them.

Table 11: Relative relationships for the cultural assistants 1 and 2, years 1998 and 1999.

1998	cultural assistant2	cultural assistant1
cultural manager	0.9405	1.0761
cultural assistant2		**(3)1.1433**
social manager2	0.7532	0.8619
financial manager	0.7683	0.9970
projects assistant	(1)1.0851	(1)1.2204
president's assistant	0.7693	(2)1.1585
social manager1	0.7495	0.8468
receptionist	0.7870	0.8239
science assistant	(2)1.0363	0.9221
president	0.8235	0.9660
director	0.9108	1.0965
conservator	0.7446	0.9861
director's assistant	0.8419	1.0016
cultural assistant1	**(3)0.9862**	
1999	cultural assistant2	cultural assistant1
cultural manager	1.082	(1)1.096
cultural assistant2		**(3)1.052**
social manager	1.036	0.844
financial manager	0.971	0.982
projects' assistant	(3)1.180	(2)1.063
president's assistant	0.767	1.028
receptionist1	1.009	0.824
science assistant	(1)1.345	0.951
president	1.051	0.940
director	(2)1.269	0.951
director's assistant	1.051	1.001
cultural assistant1	**(4)1.155**	
cultural assistant3	0.843	0.977

To the cultural assistant 1, her relationship with the cultural assistant 2 is in third place during 1998 and 1999. To the cultural assistant 2 their relationship is in third and fourth place respectively amongst 12 staff members. The results indicate that to both their connection was important. As has already been discussed, the work of the cultural

assistant 2 was divided between two areas. This is well reflected in the relative matrix that shows that above her relationship to the cultural assistant 1, the strongest ties are with the director, the science assistant, and the projects assistant. Results indicate that she worked more closely with the director during 1999 than during 1998.

TENSIONS

2003 was the last year in which the foundation issued calls for grants. It was a particularly eventful year as many people left the organization and most changed their roles and worked towards the orderly closure of the institution. The graphic in Figure 10 below shows many of the challenges involved in this method. It was produced after processing 3,727 records belonging to 16 staff members. The calculated threshold is 0.033.

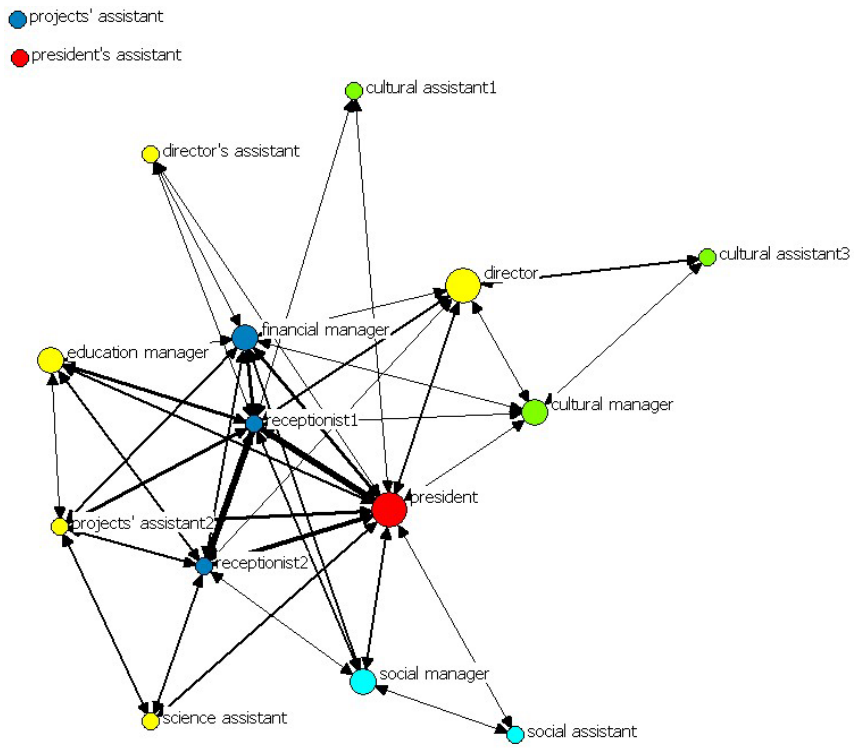


Figure 11: Network diagram, year 2003.

We can observe that the president and the receptionists are closely connected; that the cultural assistant 1 is only connected to them; and that the director has lost his ties to his area team members. Do these results mean that the receptionists worked more closely with the president than with anybody else in the foundation? or that the cultural assistant 1 only worked with the receptionist and the president? To remain updated about what happened in the different areas, the president kept final electronic versions of all the official documents produced in the organization. In turn, the receptionists kept a small number of records, a big proportion of which were the calls for grants for all the areas, which they had been sending through email to people requesting information since the year 2001.[188] These documents, containing the foundation's policies and describing the

different grants, were shared with employees from the different areas and especially with the president, whose collection containing final electronic versions of all the official documentation in the organization was also relatively small and homogeneous.[189]

Compared to the rest of the staff members, the cultural assistant 1 always kept a large number of records that included everything that she generated in relation to the organization—including calls for grants for her area and fragments of board meeting minutes and agendas—and also what she received from people outside the organization. In May her supervisor, the cultural manager, with whom the diagrams showed an above average connection for the years before, left the organization.[190] The role of the executive director also changed that year, his last one in his function, because there were no new grants and projects to implement for the future. His personal assistant left at the end of the year and the education manager supervised the ongoing grants of the area. As always he worked on the editing of books about art and cultural patrimony, this time with the help of the cultural assistant 3. His detachment from his team members indicates the changes in his role.

Represented as text mining results, the four stories are the consequence of the tensions between quantity of records and similarity. Figure 12 below explains this tension. It is a cluster analysis of a random sample of 100 records from the year 2003 belonging to the receptionist 1, the president, the cultural assistant 1 and the director.

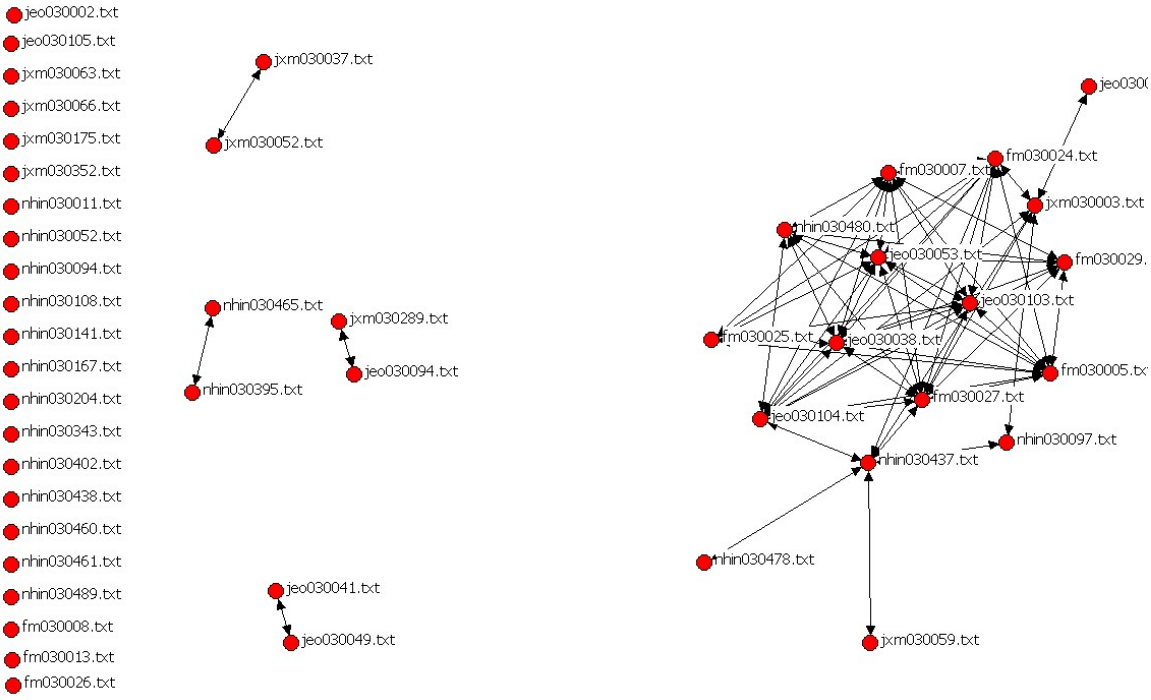


Figure 12: Cluster analysis of a sample of 100 records, year 2003.

The majority of the records in the main cluster are calls for grants and board meeting minutes; they belong to the president and the receptionist 1. Because these records are similar and/or share common pieces of text and vocabulary, they are related to each other. On the other hand, most of the records of the cultural assistant and the director are varied in content. Included among the director's records are biographies of a painter and a paintings glossary, and amongst those of the cultural assistant are movie scripts and a vita of a film maker. The vocabulary in these records is diverse and their similarity with other records in the set is low. Staff members with smaller, homogeneous collections have stronger ties, while staff members with heterogeneous and large collections are more isolated. In the organization-wide matrix, the strong relationships set the threshold averages high.

The receptionists as outlets of the foundation's calls for grants, and the president, overseeing the organization, were informed about all the programs and therefore connected to everyone else in the organization. That the cultural assistant is isolated shows that she was focused on the issues of her area and communicated with people outside the institution.[191] The director in turn was letting go his former executive functions. Consistently, these tensions are caused by the aggregation of the formal documents in the organization, such as the calls for grants and board meeting minutes, which in turn is reasonable as these are the types of records that a majority of staff members shared and the product of the institution's collective work. A complementary way of analyzing the phenomena is by filtering out the records that are producing the tension.

Sieving

Sieving refers to removing the records of one or more staff members from one or various yearly samples. This changes the statistics of the bag of words representation, the number of relationships that the staff members have in the yearly matrix, and the structure of the network diagram. To experiment with this technique I removed the records of the receptionists, the financial manager, and the projects assistant 2. All of these are small and homogeneous bodies of records that share many of the same records among themselves, with the president, and with the rest of the staff members. The idea behind this technique is to observe how removing these records affect the rendering of work relationships between the rest of the staff members. [192]

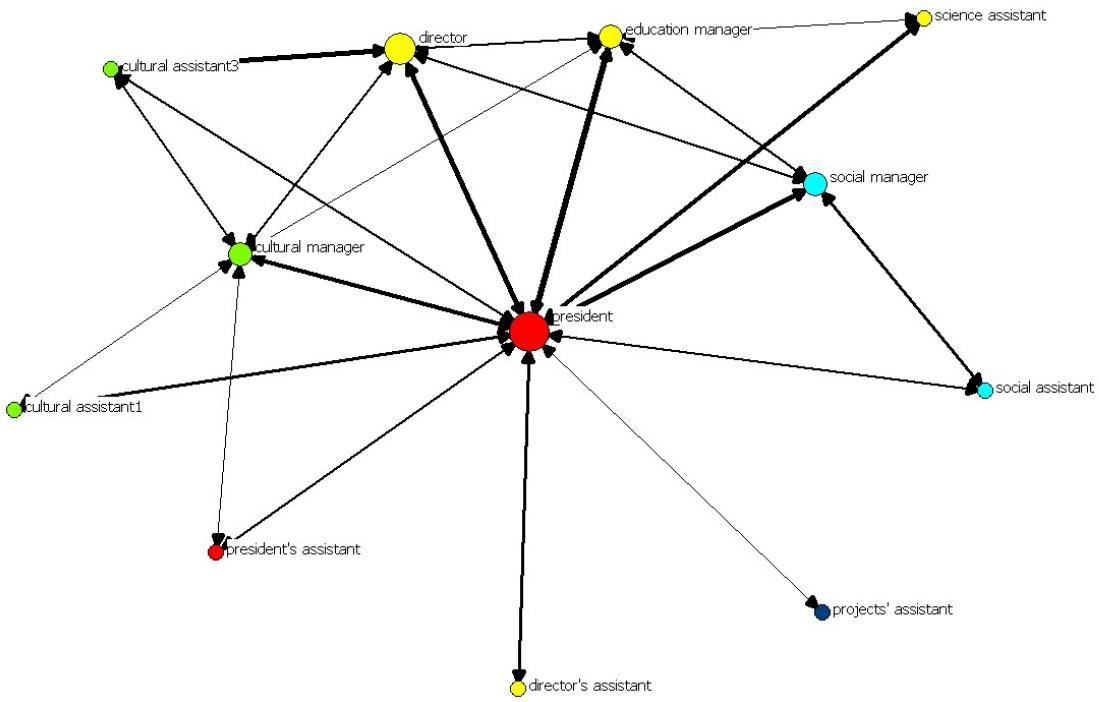


Figure 13: Network diagram of the year 2003 with three staff members removed from the set.

In the filtered diagram above the president is at the center of the organization. The director is related to those with whom he worked most including the educational manager who oversaw the remaining projects in the science and education area. In addition, his editorial work and having to step in and manage ongoing cultural projects kept him close to the cultural assistant 3. Continuing with the same pattern, the cultural assistant 1 is in the outskirts of the network with ties to her manager and to the president. Through the years, the social manager consistently maintained ties with her assistant, the director, and the president. Both views, the complete and the filtered one, are complementary and do not contradict each other. The first one shows how changes in record-making and record-keeping habits point to changes in functions and roles. The second one is consistent with

regular work relationships in the organization. Both can be contrasted with the drawing in Figure 14 from the next section *Outskirts and Endings*, which shows the receptionists in the outskirts of the network when their function distributing information about the organization by email had ended.

INDIVIDUAL RELATIONSHIPS

A question that emerged from the analysis of the diagrams from 2003—both the complete and the sieved one—was the increase in the distance between the director and his assistant, given that since 1996 and until 2003 they had maintained an above average relationship.[193] As I stated above, 2003 was a turning point for the organization which shifted from being active in its purpose to being dismantled. That year the director supervised the last call for grants and his assistant left at the end of December. To verify that the averages reflect the change in this relationship accurately I used a validation method based on the Gini coefficient, an economic measure that compares distribution of inequality or wealth.[194]

In this case, the curves resulting from plotting the cosine similarities involved in a given relationship for two consecutive years show the cosine similarities between records of the director and his assistant that ranked at the same percentage. The results obtained should match the average of cosine similarities results in the network diagrams as a comparison of one year to the next. Figure 13 below shows a graphic in which the curves obtained by plotting the cosine similarities between the records of the director and his assistant in 2002 and 2003. The Y axis shows the possible range of cosine similarity from 0 to 1 and the X axis the ranking of similarities that go from 0% (minimum similarity) to 100% (maximum similarity). The results show that the highest similarities between their

records correspond to the year 2002. This agrees with the results of the network diagrams for this relationship during those years (See Appendix III Network Diagrams).

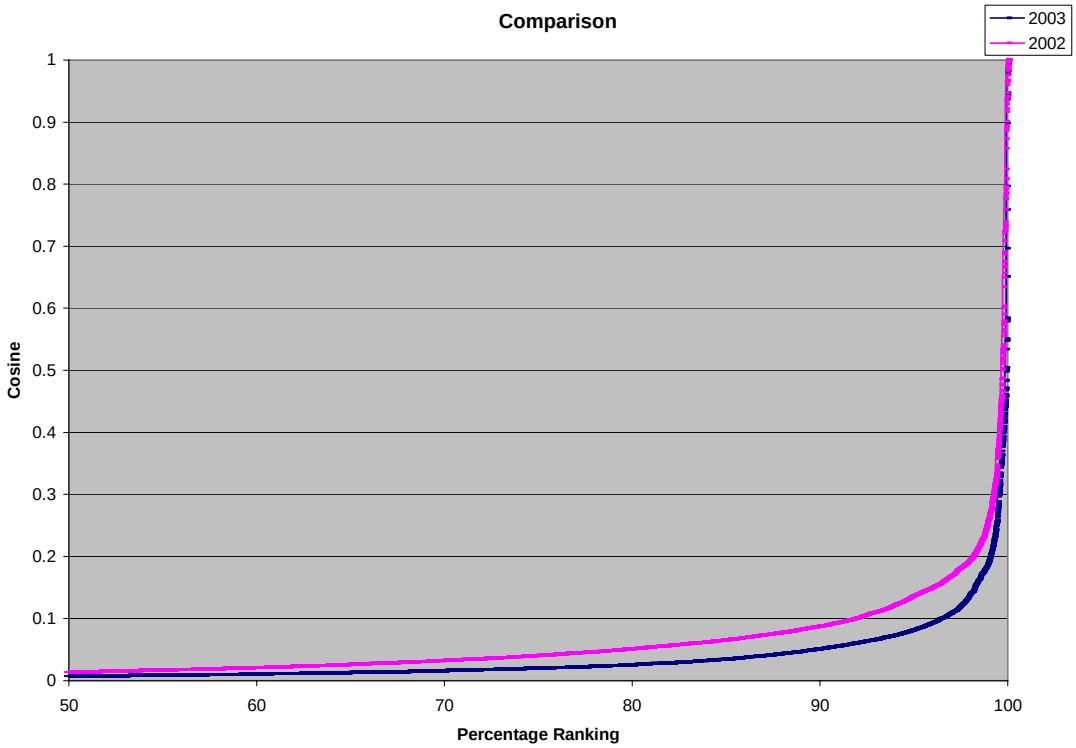


Figure 14: Comparison of cosine similarity curves between the director and his assistant, years 2002 and 2003.

This method was introduced almost at the end of this research and has yet to be applied to relationships across the matrices. Its main advantage is that it allows verifying individual average results, overcoming the fact that each one of these relationships is uneven (include different amount and content of records) from one year to the next.

Another way of looking at cosine similarities between the two staff members is to do a distribution of similarities such as the one shown in Figure 15 below. In this case, the Y axis shows the number of documents and the X axis the percentages at which

cosine similarities rank from non-similar to very similar. For each category of similarity, the cosine similarities between the director and his assistant are plotted for the years 2002 and 2003. The principle of this distribution is similar to the curve shown in Figure 14 above, except that percentage ranking of similarities can be observed in relation to number of documents.

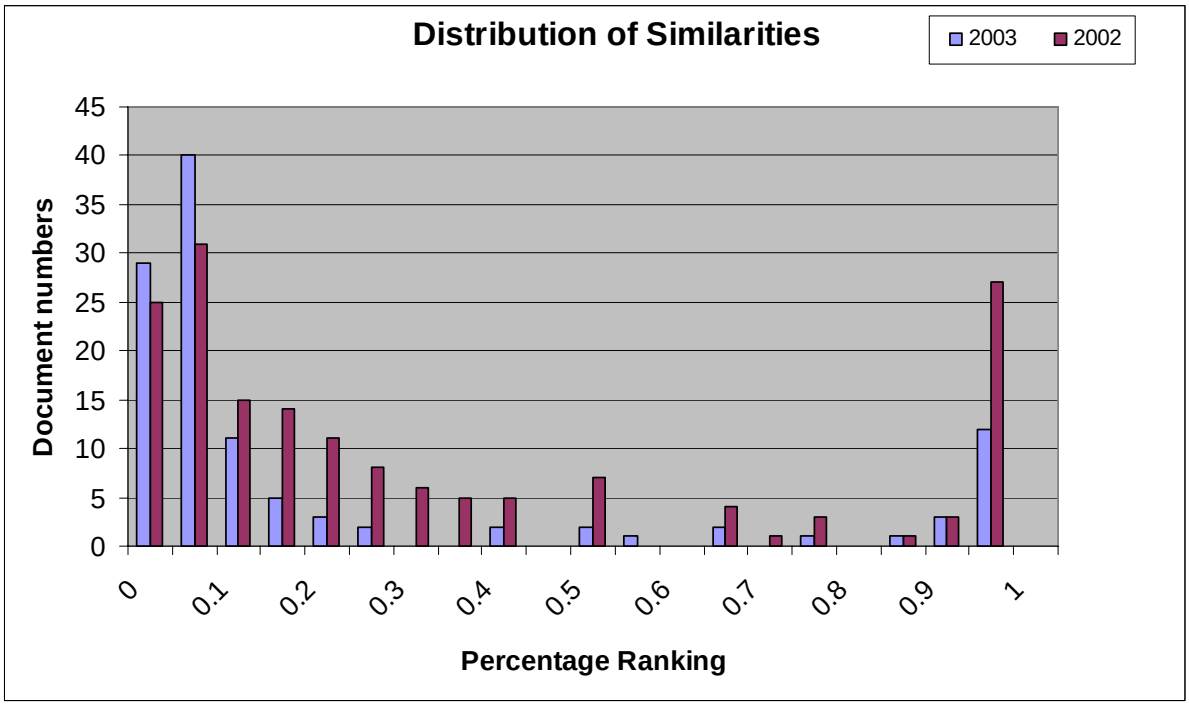


Figure 15: Distribution of cosine similarities between the records of the director and his assistant, years 2002 and 2003.

The results show that in 2003 the director and his secretary had less similarity between their records than in 2002. In other words, they worked less closely in 2003.

OUTSKIRTS AND ENDINGS

The following diagrams show cases in which staff members are on the outskirts of the network, either because they are leaving or have left the organization during a given year, or because their functions have changed. Again, underlying the results is the tension between quantities of records and strength of similarity, but different from what was described above in which the cases presented small quantities of records with strong similarity (and therefore were pulled to the center of the network), here the relation is inverse: small amounts of records bearing little similarity with the records of the rest of the staff members. In evaluating this set it has to be considered that by 2004 and 2005 most staff members were working on closing the institution. The frequency of the board meetings had decreased and there where no new plans or projects to approve, which impacted the types and quantities of records generated in the organization.

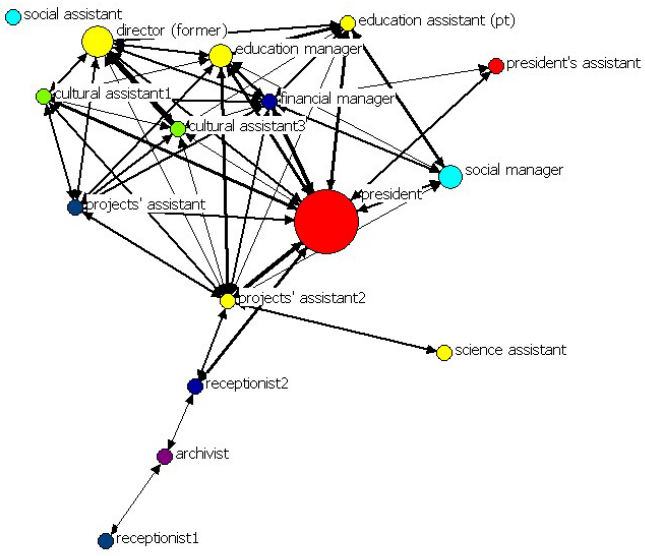


Figure 16: Network diagram, year 2004.

During the three months that I spent working at the foundation in the winter of 2004 I created very few records: requests for budgets, memos to the lawyers, various types of records inventory spreadsheets, and training materials. Among the latter was the manual with instructions to inventory and re-house paper records. Before leaving for Austin at the end of August I asked the social manager if she would coordinate the archiving project, and in turn she assigned the two receptionists to process the projects files. More than any other staff member in the organization (who auto-archived their own records but also had other things to do to close the foundation) they devoted their time to archiving. Since there were no more calls for grants to send by email, apart from the manual of procedures that they both kept in their virtual folders, they had very few electronic records.

As opposed to their central position the year before, both receptionists are dangling from the network. Each of them has few records but this time these are different from others in the set, related to their work as receptionists and to the archiving project that started that year. In fact, they are related to me because the three of us share the archiving manual that I prepared and left in my folder in the shared directory. The case of the science assistant is the same. Her virtual folder that year has few records, most of which are specifically related to her work overseeing the results of the calls for grants in the science area. After working for a few months that year she left the organization, first on maternity leave and then for good, and was replaced by the part-time assistant 2 who generated many records assisting both the cultural area and the science and education areas.

In 2005 the foundation moved to a smaller office. The receptionist 2 who had left in May and the receptionist 1 stayed part-time until the end of the year to close the last

project files. The education manager was the last manager to leave the organization; she supervised the last projects and with the help of the projects assistant aided the president overseeing different aspects of closing the organization. The social assistant also left in May. Her close relationship to the projects assistant is functional, as they both sent letters to grant recipients reminding them about the foundation's closing date and asking them for final reports, or responding to last minute petitions. The network drawing below reflects the last year of the foundation, with the president at the center of the network, the education manager closer and with ties to most of the personnel, and the receptionists and most everybody else on their way out of the institution.

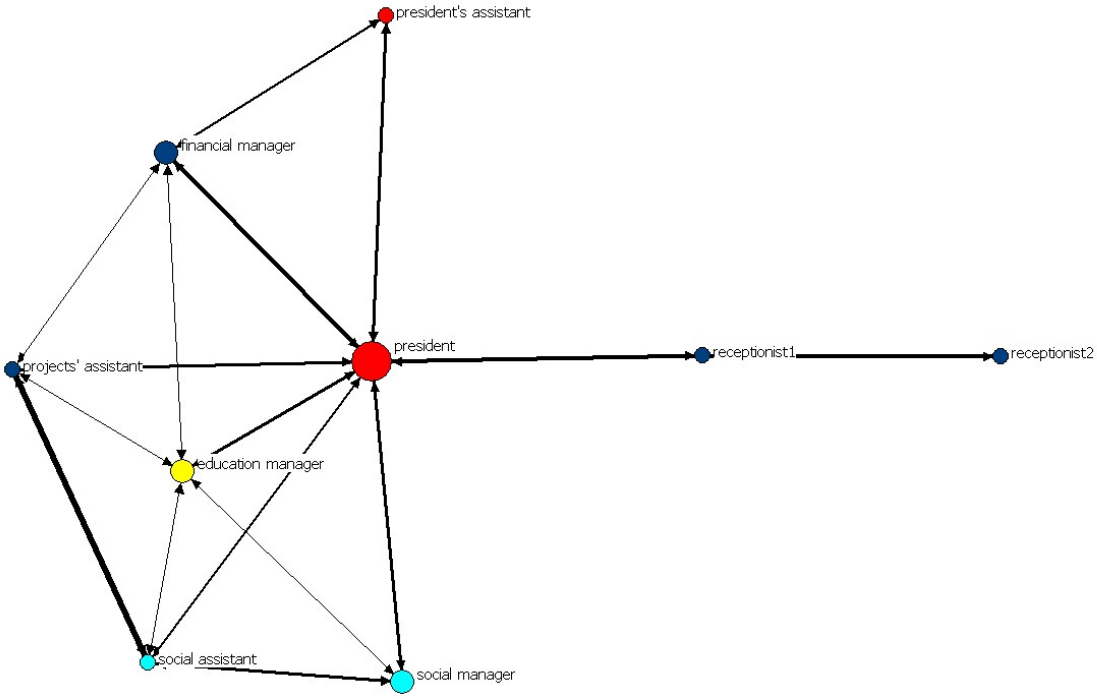


Figure 17: Network diagram, year 2005.

These results show that in the influence on the equation of quantity of records and similarity, the proportion of records similar to those of other staff members is what determines the strength of their relationships as seen through this method.

Digital Archival Appraisal

My intention in developing this appraisal method is theoretical. The goal was to answer the question of whether the records of the natural archive—focusing on the text records—would portray the organization that created and maintained them and thus could constitute evidence. For this reason I used a bulk approach and included all the text records in Spanish in the shared directory belonging to all the staff members, even those who worked part-time or who only worked for a number of months in any given year. The principle of the appraisal method is that by using text mining as an independent analysis method, the relationships between texts and therefore between their creators will emerge inductively. In the process of interpretation, the use of social network analysis provides both a theoretical framework and a visualization aid for organization-wide relationships among staff members. The use of animation presents the data from the point of view of one staff member and the evolution of his relationships through the years.

For interpretation purposes, job functions, events, dates, programs, record-keeping and record-making practices, work practices and changes in roles are taken into consideration. Finally, results are validated through the results of the interviews, observations of quantities and contents of records, the narrative of the archive's formation process, and statistical distributions that allow checking whether the inferences made about the relations, based on the text mining results, are sound. In all cases we have to bear in mind that these relationships are based on text records; that some people wrote more than others, kept more than others, or deleted more than others, and that "relationships" are contingent upon the number and types of records involved. And yet, the unbalance emerging from the analysis not only is due to gaps and repetition in the

archive, but also seems to reflect the changes that occur in any person's work history and in any organization.

Results suggest that using this appraisal method with the records of a natural archive, it is possible to resurface aspects of the organization's work dynamics as they occurred in the past. Predominantly the sets provide snapshots of the organization's structure including functions and staff members' roles that go beyond the organizational chart and the job descriptions. Figure 18 below presents the organizational chart of the year 2002 and the network diagram of that year, showing the two representations of the organization, the first one formal and idealized and the second one inductive.

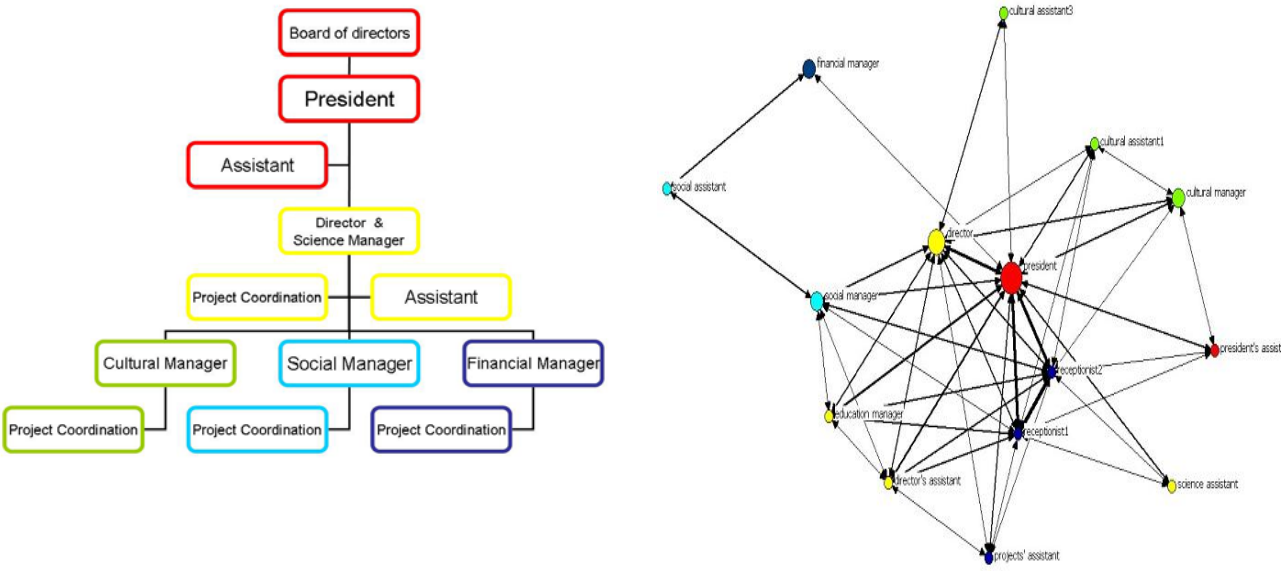


Figure 18: Aleph's organizational chart and network drawing, year 2002.

Between the two graphics there are agreements and contradictions. While the basic official structure of the organization is reflected in both (the president at the center tied to the directors and managers), aspects of lateral collaboration between areas and

people are best reflected in the network diagram. Among the latter, the tie between the cultural assistant 3 and the director (both involved in the production of a catalogue), the role of the receptionists as the institution's external communicators, or the connection between the social assistant and the financial manager (the major part of her job was requesting payments for the area projects), are examples of how certain aspects of work dynamics are reflected through the records. But results are not always clear-cut and there are many elements such as number and types of records involved and individual record-keeping practices to consider during their interpretation. Indeed, I have showed the way in which I had to deal with certain ambiguities present in the results and pointed to those that need to be clarified.

This appraisal method highlights the different roles that records play. Depending on their provenance—that is on where records are located in the shared directory—a record has a different function that in turn reflects the role of its holder or creator. The call for grants found in the receptionists' folders indicate their role distributing information about the organization, but in the folder of the area manager it points to the creator of the program addressed by the call. Because records in the natural archive reflect work processes, analyzed in aggregate and over time, they reflect work dynamics. This suggests that if this archive were appraised for selection with the goal of discarding repetitive versions, the roles of many of these record holders would be erased.

Appraisal considerations are not divorced from evaluating how to make sense of large quantities of records. With paper records, appraisal is constrained by quantity and by the resources involved in processing them. With digital records the issue at stake is apparent chaos and how to approach it intelligibly without disturbing its original order. This appraisal method allowed making sense of 16,000 records, focusing on the way in

which people in the organization worked with each other over time. Moreover, as long as the digital records' file structure is maintained and documented to preserve aspects of provenance and original order, there is potential to use other data-driven methods, such as clustering and machine learning classification, to structure digital records for multiple purposes and from different perspectives.[195]

So far I have appraised the archive including all the records (albeit under the restriction of Spanish language selection). However, I have not answered the question of whether the description of the organization would have been more or perhaps less accurate had the electronic archive been weeded a priori as is common archival practice. Certainly, due to the nature of the method, the structure of the diagrams would change, but it is still to be seen whether and which relationships would change and how. Now that I have the complete image of the archive, the next step would be to create other distributions of the archive and submit them to this method to test other hypotheses. For example, I can use macro-appraisal or the guidelines from the National Archives and Records Administration (NARA) to select records and compare the results of the different representations, for example, find out what will happen with the relationships between staff members if all the versions of the board meeting minutes are removed from all the folders except from that one of the president who was chair of the Board of Directors.

As it is, this method stresses the completeness of the archive; each record, either because it is very similar to or very different from others and depending on its location, says something about the organization. It also shows that archival bond, understood as different levels of similarity between records, can provide the structure—in this case relationships between staff members and functions—that the natural archive does not

have to begin with, and that as done in HCI and IR projects, could be useful to make sense of the archive and to create points of access.[196]

From the beginning of this project I thought how I would have validated the results if I did not know anything about the organization, could not talk to anybody about their roles and record-keeping practices, or if the archive had lost its original structure. Future work will help establish whether we can trust the method in archives for which we have little or no information about their formation process. Indeed, without the interview data it is not possible to state that so and so worked together or not. In principle however, the vector space model/Tf-idf/cosine similarities/ is a reliable (and thoroughly tested) method to determine similarities between records. Coupled with statistical distributions of cosine similarities it is plausible to identify who worked closer to or distant from whom according to their records and those results could be used as starting points to understand an organization. At this point, this method is still a basic road map, one that opens the door to the use of digital tools to appraise digital archives.

APPRAISAL RESEARCH FOLLOW-UP

In order to improve the appraisal method, complement and validate it, or to use the lessons learned from it, other research projects must be pursued. At issue is the extent to which relationships are based on the use of common words. The implementation of Tf-idf attempts to highlight unique terms in the midst of repetition. Given that each set and each relationship is different from every other, a way to reduce or eliminate ambiguity from the sets will be to apply feature extraction (removal of high frequency words) in the sets, and observe whether relationships are maintained or not. Application of feature extraction will require obtaining a list of words from the virtual folders of each staff

member for each yearly set and apply them as stop words during the construction of the vector space model. I will also continue testing the internal validity of the method through analysis of the distributions of cosine similarities across all the yearly sets. Text corpora belonging to different archives differ in length, style, vocabulary, and language. I will continue testing the inductive text mining appraisal method in other digital archives (more and less structured) to establish if it is generalizable.

In terms of external validity, Dr. Michael Khoo suggested that I have the interviewees rate their relationships with the rest of the staff members in the yearly sets and to compare those results with a curve drawn of the average of cosine similarities in the matrices.[197] And yet, a specific point to be made will be to clarify to the interviewees that they have to consider their relationships from the point of view of the work and records that they shared, and not from a personal perspective. Perhaps this methodology is more appropriate to test people's perceptions about their work relationships and how they map with the digital appraisal results of their work records. Another suggestion, made by Dr. Allan Renear, was to map the visualizations to organizational behavior theories.[198]

Appropriate visualization of text mining results supports valid interpretations of the phenomena being studied. With the TACC Visualization Team I am producing an interactive visualization that reflects the changes in work dynamics between all the staff members over a period of ten years as well as the relative relationships of each of them. The software used in this case is PREFUSE.[199] We would like to build in the application the possibility to visualize the texts included in the different relationships which would help establish the nature of the activities that they shared.

The use of computer assisted tools for archival processing was identified as a need in the New Skills for the Digital Era Colloquium in May of 2005.[200] I want to extend the experience obtained using text mining techniques to explore electronic text records arrangement through clustering and classification, and archival description through feature extraction. Such projects will allow improving the flexibility, the functionality, and the documentation of the text mining program developed with Dr. Hai Bi.

PRESERVING THE ARCHIVES' REPRESENTATION

The results of the digital appraisal method indicate that the records in the natural archive have the potential to represent the work processes and dynamics of Aleph and that the other digital objects that surround them on the networked server contribute to confirm or contest this representation. Furthermore, the data-driven representation constitutes a model against which other representations can be compared and contrasted. Independently from whether the Aleph archive will ever be made publicly accessible, and from the issues that will have to be addressed if this happens, the challenge now is to preserve the structure and contents of the shared directory "as is," at a minimum during the records retention period, considering that in the future all or some of its contents may be retained for the long term. The next section presents the preservation strategy designed for the natural electronic archive to achieve that goal.[201]

PART IV: FUTURE

Preservation Framework

Digital preservation is about layers of hardware and software; records and data; standards and innovation; individual responsibilities; and institutional commitments. These layers are not autonomous from each other, they aggregate and each passes on its legacy to the next. Ideally, these layers must converge to render a digital object meaningfully over time, but time and idiosyncrasies can turn the layered structure into a Babel of wills and codes that may not interoperate.[202]A bitstream will be accessible and readable with compatible software that in turn functions within a compatible operating system on compatible hardware. Over time, the technical layers must be sustained by individuals and institutions that function as archives. A preservation strategy for Aleph's natural archive had to address all these layers and consider the financial limitations, the requirements of access and functionality during the retention period, and the uncertainty of its future.

Regarding what to preserve, the strategy had to account for the outcome of the formation process studies and consider the need to preserve all the contents of the archive as a digital archaeological site. The digital archive "as is" provides evidence of how Aleph made IT decisions, how the introduction of networked computing changed work-practices, how people created and used electronic records before they were acknowledged as such, and how the relative value of electronic and paper records changed over time. This idea is supported by the results of the appraisal method, which suggest that despite the chaos of its structure, its repetitive nature, and its unaccountable gaps, the archive still provides evidence about the organization and it is possible to make sense of the records.

Once the decision to keep the records "as is" was made, the next steps involved determining where to keep them, how to prepare them for the transition, and what needed to be accomplished for their maintenance. The strategy's core tactics are bitstream preservation, migration on demand, and virtual migration. It comprises transferring the contents of the networked server unchanged to a new server environment or "dark archive"[203] with the purposes of overcoming the shortcuts of computing equipment obsolescence. In the end, a preservation strategy that maintains the archive intact while moving it forward to allow access aims to provide a theoretical understanding of the relationship between preservation practices and electronic records creation and use.

As it has been described in Part I, *Entry in the Field*, the preservation strategy went through various iterations. From the original idea of using an IR system, to creating a living dark archive under the custody of a telecommunications agency, to the minimalist preservation strategy that is described in this section, the final form contains some legacy decisions that were better suited for previous versions of the plan. And yet, the process of negotiating and developing the strategy (all of which occurred during a period of two years), resulted in hindsight in the best possible approach, one that will allow further preservation and access strategies without modifying the original structure of the archive. The following description addresses the main elements that were taken into consideration while devising the digital preservation strategy for the natural archive.

POST CUSTODIAL CONSIDERATIONS

Economics and logistics

Aleph's archive is a case of post-custodialism, in which the creators became responsible for the custody of their archive. [204] My role was to guide the archiving

process to protect the records and make sure that the owners remain in compliance.[205]As an organization Aleph no longer exists, so the burden of custody is born by the last president and members of the board of the umbrella organization. Therefore the challenge was creating a cost effective, technically feasible, and legally sound preservation strategy spanning the entire records retention period.[206]

Aleph invested significant efforts and money in preparing their paper archives for storage. They paid ten years in advance for the services of the commercial records management company where the paper portion of the archive will remain dormant until the retention period expires and provisions are made for the records to be accessed only by authorized individuals. While the chain of custody is still unbroken, the physical care of the records rests with the company. Aleph also made a significant initial investment to preserve the electronic records, by purchasing the new server, the audit and control software and the IT consultants to transfer the records, but the problem of maintaining them over time is more complex. In devising a digital preservation strategy I had to make sure that the custody of the digital portion of the archive was not going to become a burden involving high costs, IT consultants, and coordination of activities.

The archive's view

In planning how to preserve the natural archive, the way in which it was maintained for almost twenty years was revisited. Starting from a few networked workstations with minimal storage capacity and the DOS operating system, the institution improved their computing systems with prudence and regularity.[207] During these upgrades, records and old applications were moved from one shared server to the next and subsequent operating systems, but they were never migrated to the format corresponding to the current version supported by their creating software; only the data

from the databases was migrated from one database system to the next. As a consequence, DOS compliant applications still on the server do not function properly, and DOS records from 1991 through 1996—when the foundation stopped using DOS—can be rendered only with defects. But also as a consequence, the bitstreams have been preserved essentially intact.

This suggests that at the moment of making IT upgrades, maintaining functional versions of previous database models was not a consideration. It also seems that keeping old DOS records was a matter of individual preferences or of the holder's neglecting to migrate them, and the fact that they could only be rendered with defects was not an issue for the staff members, who possibly did not ever need to open them again.[208] For the purposes of this preservation strategy, at the moment of deciding whether older records and non-functional systems had to be migrated, the view that staff members had of records and systems was taken into consideration. At this point it did not seem necessary to migrate the records to newer versions. What did seem necessary was to implement a system to follow through technology changes and anticipate the archived formats response to them.

Access and moving the natural archive forward

Attending to the doubts about what to do with the archive in the future, given the legal vacuum that exists in Argentina and emergent legal trends coming from abroad that suggest that everything digital may constitute evidence, I considered that a prudent measure was assuring access to the records and latest iterations of the databases in case information had to be produced during the records retention period. At the same time, the possibility that in the future all or some of the databases and the records may be retained for the long term had to be considered. This meant that the dark archive could not remain

like a "black box,"[209] closed and waiting to be overcome by obsolescence. Instead, the natural archive must be moved forward virtually, its evolution has to follow technology changes, and the integrity and authenticity of its contents have to be continuously monitored. This required developing routines to accomplish and document during the next ten years. Even if they prove redundant or overkill for an archive that might not be consulted at all during its retention period, these routines and procedures would assure that this group of digital objects remains an archive.

Authenticity and integrity

According to the Society of American Archivists glossary, records' authenticity is "the quality of being genuine, not a counterfeit, and free from tampering, and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context." Records' integrity refers to "the quality of being whole and unaltered through loss, tampering, or corruption."[210] In the context of an undisciplined record-keeping environment and of shared records creation practices the concepts of authenticity and integrity had to be revisited to determine exactly what they mean and to translate them into the digital preservation strategy.

Authorship, location, dates, and degrees of completeness are some of the elements used to evaluate authenticity in records. For many records in the natural archive provenance—understood as authorship or the creating department—is shared; their component fragments and versions are scattered throughout the virtual folders of the staff members across time. At Aleph, in spite of changes in personnel and in individual records organization practices, the directory structure persisted for fifteen years. In a record-keeping environment in which authorship is blurry, the location of the files in the directory—understood as file path—constitutes perhaps the best indication of provenance

because it points to their creators, co-creators, and/or gatherers. In turn, the shared directory structure including the sub-directories provides hints about original order and record-keeping rationale.[211]

Issues of integrity refer to the possibility that records might be altered, tampered with, or distorted in some way. While Aleph was an active organization, electronic records and databases could only be accessed by staff members. Getting to the shared directory required a password to the local network; a second password was needed to use the databases.[212] From the narrative we can infer that most staff members did not even look at each others' records without asking permission; however, editing, cutting, and pasting from one record to the other was an established practice. All of this indicates that the possibility that records in the shared directory were altered or tampered with was minimal, but the fact that they were co-created should remain evident. For the purposes of the digital preservation strategy I understood that integrity and authenticity were intertwined. Keeping the entire contents of the shared directory in the final structure, including all the contents as they were used, offered the possibility of understanding shared practices and preserving the available contextual and structural elements of the natural archive.

PREVENTIVE CONSERVATION

The preservation strategy developed for the natural digital archive is informed by the concept of preventive conservation. Preventive conservation comprises a variety of studies and practices whose goal is to minimize the deterioration of cultural patrimony over time.[213] Its recommendations and procedures are based on a deep understanding of the basic materials that compose different categories of cultural patrimony such as paper,

metals, wood, cotton, plastic, etc. and how these are affected by the environment and by use. Based on this knowledge, the ideal conditions in which broad categories of objects of single and composite materials should be stored, housed, exhibited, manipulated, and documented are established, making it possible to focus on the overall protection and management of collections. Preventive conservation also addresses prevention of natural and human caused disasters by applying a combination of general security measures and others specific to the environmental and social context in which the cultural materials are located. By minimizing the cultural objects' deterioration risks and the prospective need to conduct interventions, preventive conservation seeks to assure the protection of the objects' authenticity and integrity. This approach also has economic implications as it allows caring for large numbers of items which, in the long run, is efficient and cost effective.

In the digital preservation strategy I apply preventive conservation principles by: understanding the archive's formation process and the technical characteristics and dependencies of its components; providing a technical environment in which the majority of the different types of digital objects can be accessed; controlling access to the archive; establishing security and risk prevention measures; performing continuous testing of file rendering and database functionalities in updated technical environments to assure future access; and generating continuous documentation of the use and behavior of objects to allow further decision making. The strategy was devised throughout the years during which I worked on the archiving project and, as time passed, circumstances and decisions changed. Therefore there are vestiges of legacy decisions in the strategy and I will point to them as I explain the reasoning behind and the components of the minimalist preservation strategy.

Minimalist Preservation Strategy

The preservation strategy devised for the natural archive aims to:

- Maintain evidence of the archive's formation process
- Maintain records and systems accessible, at a minimum during the required retention period
- Maintain the integrity and authenticity of the records and systems over time
- Demand only minimum computing and maintenance costs
- Continue the documentation of the evolution of the archive

To accomplish these goals all the digital contents as they existed in the networked server were transferred to a dark archive. Once in the dark archive, a combination of tools and procedures assures that the archival contents preserve their authenticity and integrity while remaining accessible. Foreseeing limited use, to increase security and to avoid maintenance overhead, the dark archive will be kept off-line and un-plugged. It will be turned on only by authorized users to retrieve information when and if needed, and to perform server maintenance routines and records integrity checks. Backup copies of all the contents of the dark archive have been made, and in case of failure of the dark archive, the records will be restored from this backup. One of the backup DVD copies is used to perform migration tests and to decide whether the software, the hardware, and the records or systems need to be migrated in order to provide access. Independently from deciding whether or not to change hardware and software during the next ten years, these tests will enable making sound decisions when and if it becomes necessary to migrate the technical environment or the electronic records.

Both the transfer steps, including the technical specifications to prepare the dark archive's environment, and the maintenance procedures are detailed in the Transfer and

Maintenance Protocol, written between June of 2005 and November of 2006, and included as Appendix IV. This document was informed by the results of preliminary transfer and migration tests, by archival acquisition best practices, and institutional repository ingest practices learned in my courses in the School of Information at the University of Texas at Austin, and on recommendations for electronic records transfer and retention included in the section Retention and Review of the publication Preservation Management of Digital Materials: The Handbook.[214] The Transfer and Maintenance Protocol outlines the preservation strategy devised for the digital archive throughout governance by the records retention schedule. It was prepared for the IT consultant to follow during transfer and maintenance of the records to the dark archive, to inform the foundation's authorities about the procedures followed, and as documentation that attests that best practices were and will be followed to support the authenticity of the records. Regarding the networked server, we planned to keep it at least until the accuracy of the transfer could be verified. [215]

The dark archive was set up during June and July of 2006. It was purchased with the operating system in place, and most of the rest of the set up steps (rendering software installation, anti-virus installation, and transfer tests) were done by me together with technicians from the IT consulting firm that had worked for Aleph during the previous ten years. The official transfer into the new server was done in February of 2007 following most, but not all of the steps indicated in the protocol. A number of lessons were learned from the experience: some protocol steps that I should have enforced, others that I should have researched more closely, and others that I missed. The following sections present the decisions, procedures, and lessons learned from the minimalist preservation strategy.

.

DARK ARCHIVE

According to data gathered from the accounting books and recorded in the Metadata Timeline (See Appendix II) the existing networked server, a Hewlett Packard NetServer LC3 with Windows NT OS, had been purchased in 2000 and a previous one in 1996, indicating a server turnover of five years.[216] To that fact must be added the considerations that disk failure is the most frequent cause of data loss [217] and that computer fan bearings have a limited life span (five to seven years). These considerations urged us to purchase a new server system to function as a dark archive. By taking these steps, we expected to avoid mechanical problems during the next ten years. This preventive measure was also convenient from financial and practical perspectives, as it was easier to purchase a significant piece of equipment while the foundation was still functioning. Besides the goal of synchronizing the new hardware with new software, the dark archive's upgraded platform raised the issue of evaluating how files and databases would render and function in the new technical environment. Preliminary testing conducted in 2005 and detailed in the *Test* section below indicated that files and databases would remain accessible despite migrating the technical environment, so in July of 2006 the new server was acquired.

The new Hewlett Packard Proliant ML 350 server with Windows 2003 OS was designated as ArchivoDigital (Digital Archive); it was configured with RAID 5 storage technology to improve data security through storage redundancy, and a unique administrative password was assigned to it.[218] The logical partition disk on which the contents of the shared drive would be stored was configured as an NTFS file system. To access most files present in the archive, current versions of compatible applications and

file viewers were installed in the installation directory.[219] These are: Microsoft Office 2003, which at the time was the latest version available, a file management program, anti-virus software, file viewers, text editors, and the National Library of New Zealand metadata extractor that I considered could be useful in the future. To load the programs that were available online, the server was momentarily connected to the Internet through the local network.

An audit and control program, Tripwire for Windows, was installed to perform file integrity checks through MD5 hashes and to monitor access to the server.[220] MD5 hashes are generated through algorithms that convert a given file into a unique digital stamp that is irreversible and unrepeatable. Tripwire produces initial MD5 hashes of all the files in the archive. These are later checked against new hashes every time the software is run to control if files had been altered. This measure assesses both the authenticity and the integrity of the records because it confirms whether the record continues to be what it originally was and because it detects if the object has been tampered with.[221] Also, the software detects access and actions in the server including who, when, and what.

Other directories in the dark archive, for which I devised specific naming conventions, were assigned to store software applications that were not shared, such as the scheduling and the human resources software originally installed on local computers when the organization was active. A documentation directory was created to keep all the documents generated throughout the transfer and maintenance processes. Those records include the Transfer and Maintenance Protocol, the pre- and post-transfer inventories, and the documentation resultant from the transfer; the yearly logs generated by Tripwire will be automatically directed to sub-directories (named by year) in this directory. For

security reasons the server will be maintained off line, and to avoid maintenance overhead it will be maintained un-plugged.[222]

TESTS

Transfer and integrity

In June of 2006, the IT consultant and I performed a preliminary transfer test. The dark archive was connected to the local network and configured so that data could be copied from other computers but not removed from them. The transfer test was done using Total Commander 5.11, a file management program for Windows that uses file transfer protocol (FTP) and file comparison by content, and with which the consultants had experience and wanted to use for the actual transfer.[223] The test's purpose was to help me learn whether the file management program was adequate for the task and to help me define the steps that would constitute the transfer protocol.

Using the file management program, we copied a group of records and verified that they had been transferred without error. Total Commander 5.11 compares the contents of both directories bitwise and highlights with color the file that was not transferred or that is not equal to the one from which it was copied. The second test involved eliminating a file from the dark archive and running the file comparison function; the program highlighted the missing file in the networked server. In this way we became confident that the file management program would perform an accurate transfer and flag any errors if they occurred. By making sure that a complete copy of the files is transferred to the dark archive, the integrity (and authenticity) of the archive is assured.

We noticed that the file management program did not produce a log of the transfer, but since the directory trees from both servers (networked server and dark

archive) appeared on the screen, we agreed that when the actual transfer was done, a snapshot of the screen could be saved in place of a report. This recommendation proved impractical. As reported by the technician who accomplished the official transfer, documenting the results of the entire directory tree would have implied taking more than 500 screen-shots. Fortunately, I had planned to require pre and post transfer inventories which would attest to the completeness of the transfer. [224]

File rendering

Another consideration was determining whether copying the archive to a new technical environment would affect rendering of the files or the functionality of the databases. In 2005, to survey the contents of the networked server I connected my laptop running Windows XP to Aleph's network and opened the records with my local software, Microsoft Office Suite 2003. Most of the random texts, spreadsheets and presentations pulled from various folders in the shared directory rendered correctly in the updated environment of my computer, and only a few early records rendered with glitches but were still legible. I also created a path to the grant tracking system written in Clarion 1.0, and was able to use it without problems.[225] This information was recorded in the server survey that I completed as I was making the observations.

In July of 2006, prior to the official transfer, I decided to conduct a systematic file rendering test of a group of Word records, one corresponding to each year from 1991 to 2004 from the director's virtual folder.[226] The test consisted in opening the same record in three different computers running different technical environments: the office work-stations equipped with Microsoft Office 97 for Windows 98; my computer running Windows XP and Microsoft Office 2003; and the dark archive prepared with Windows Server 2003 and Microsoft Office 2003. The dark archive was not connected to the local

network. The results are shown in the File Rendering Testing Table included as Appendix V and interpreted below.

Rendering of files from 1991, 1992, 1993, 1994, and 1995 created with Microsoft Word versions 5.0 and 5.5 for DOS is equally defective when opened in any of the three environments. In all cases, Word prompts a file converter from which the user can choose to "make the file readable" with Windows or MS-DOS text encoding. Figure 19 below shows the prompt and the options.

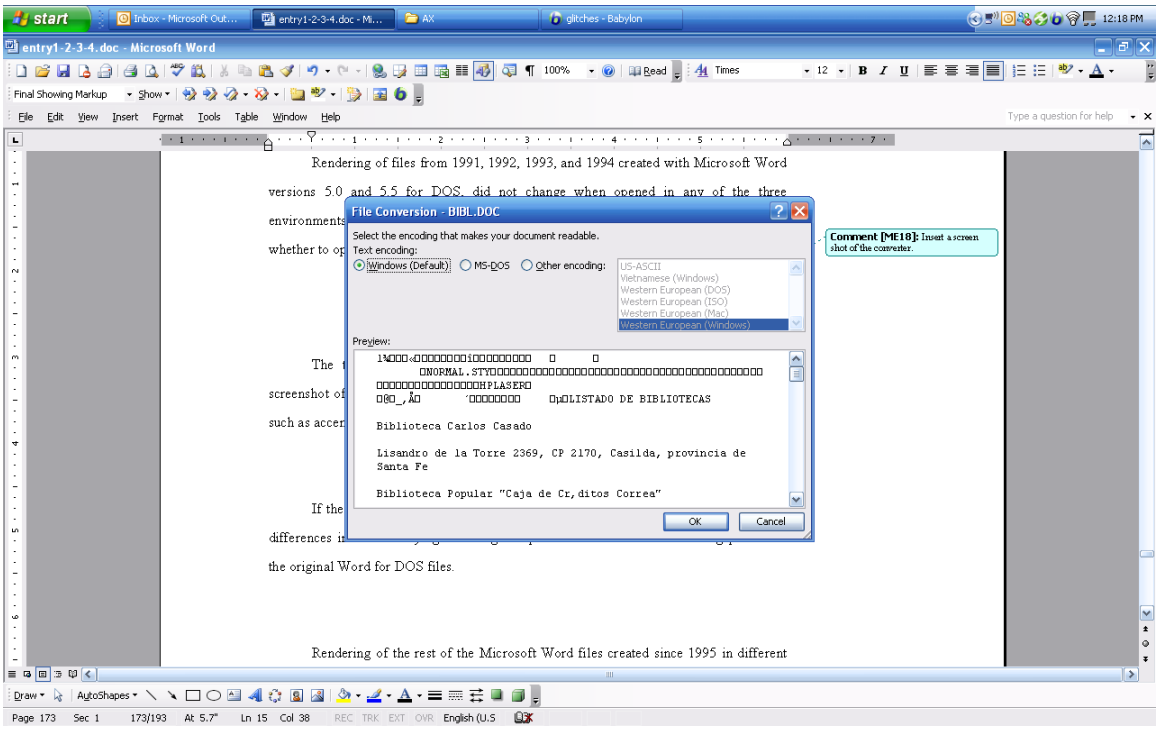


Figure 19: Screen shot of the file conversion prompted by different versions of Word for Windows to open Word for DOS files.

The types of glitches present in each of the files will depend on the encoding selected. The screen shot below shows a file created in 1992 with Word 5.5 for DOS and

opened with Windows encoding. It can be observed that the Spanish diacritics such as accents render defectively, as well as some control codes.[227]

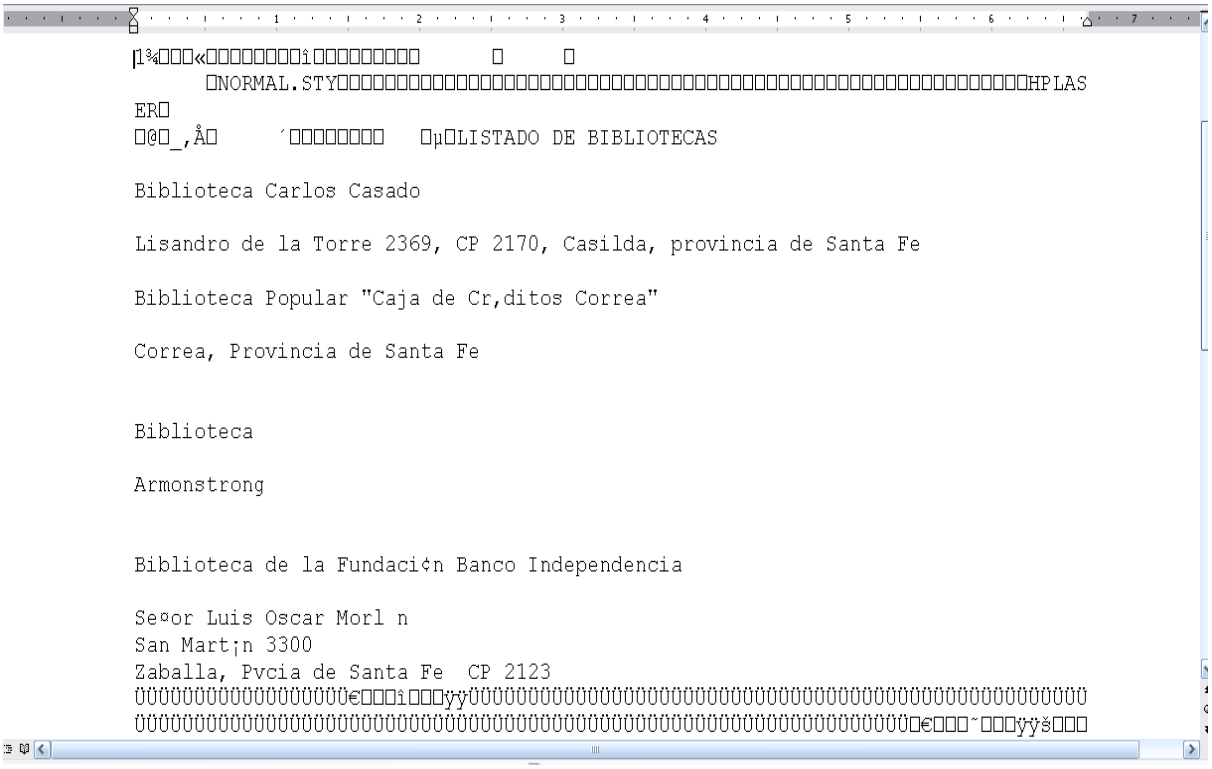


Figure 20: MS-DOS Word 5.5 file rendered with Windows text encoding.

If instead MS-DOS encoding is selected, as shown in Figure 19 below, the Spanish diacritics render correctly and there are only differences in the rendering of certain control codes.[228]

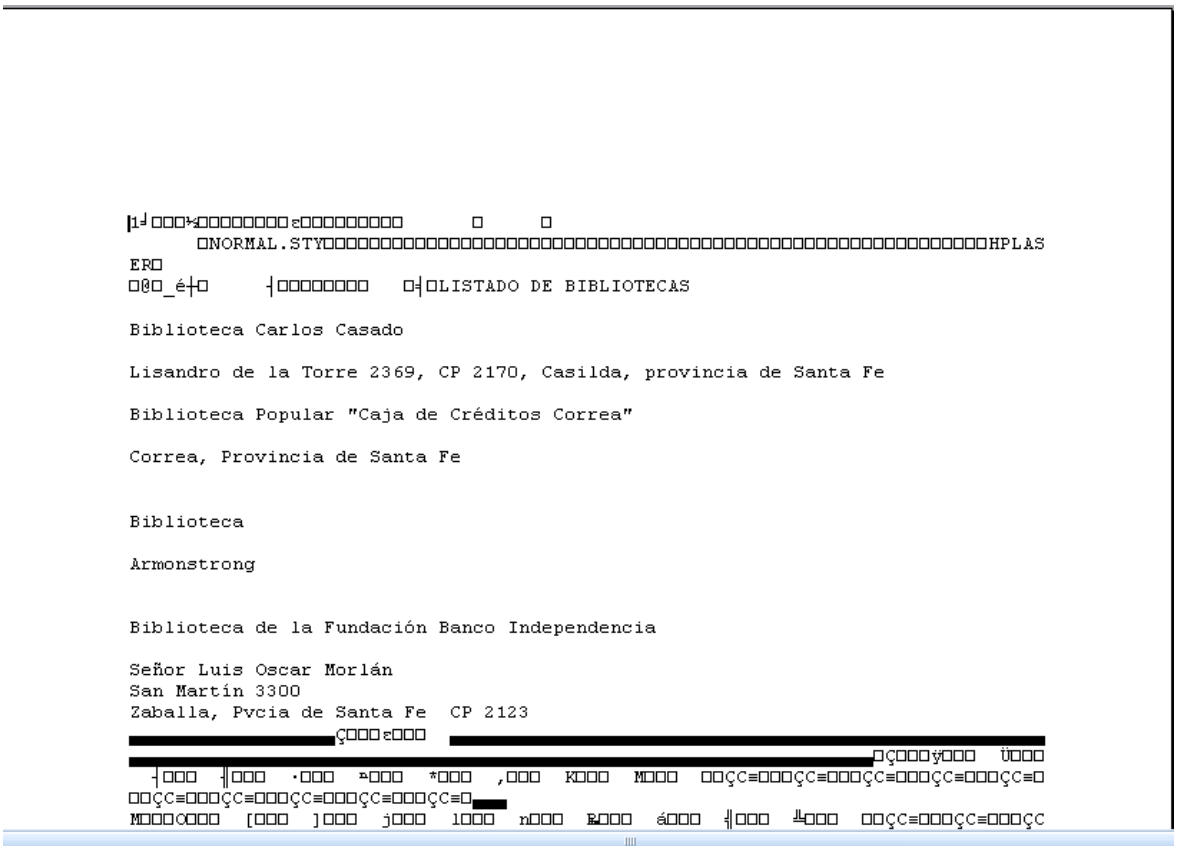


Figure 21: MS-DOS Word 5.5 file rendered with MS-DOS text encoding.

Finally, I used FileMerlin to transform the MS-DOS Word 5.5 to the current version of Microsoft Word for Windows 2003. The result is shown below. Formatting glitches are gone and Spanish accents appear correctly, as FileMerlin specifically supports MS Word versions 1.0 to 5.5 (DOS).

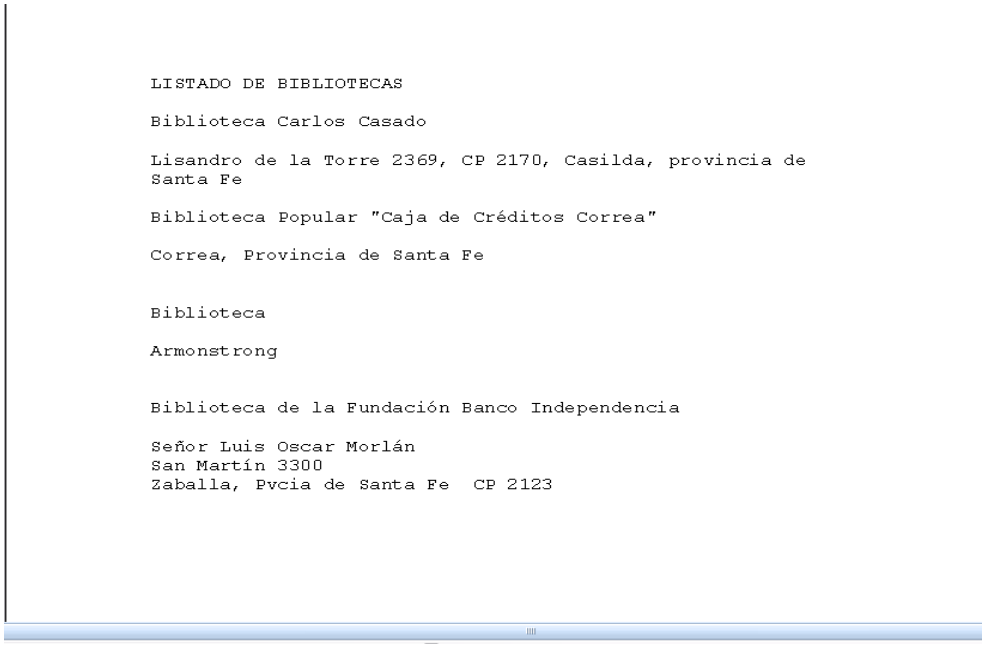


Figure 22: MS-DOS Word 5.5 file migrated with FileMerlin to Microsoft Word 97 for Windows.

Rendering of the rest of the Microsoft Word files, created since 1995 in different versions for Windows, did not present problems. These results indicate that the way in which records were available to the users has not changed since the institution moved to a Windows environment thirteen years ago. It was decided that migration of files created before 1995 would not be made, as files might not be accessed at all during the records retention period and because despite minor glitches, records are still legible.[229] If migration had been decided, only the group of files belonging to the DOS era will have to be considered for migration. If in the future migration is carried out, I will not use or change the original files but will make a migrated use copy.

TRANSFER

On January 5 of 2007, when the foundation left its office space, the transfer/copy of all the contents from the networked server to the dark archive was carried out by the IT consultants over the local network (See Appendix IV Transfer and Maintenance Protocol). Transfer steps included: automatic inventorying of the contents of the old server (pre-transfer) and those of the dark archive (post-transfer). Upon transfer, a file comparison was run with Total Commander as well as a virus check of the incoming contents. Following, Tripwire was used to create MD5 hashes of all the records stored in the ArchivoAleph directory. These hashes are saved in a database and used as controls to perform future file integrity checks. During the transfer process, the following documentation was produced:

- A report including a list of the software installed in the dark archive and the name of the directory in which they are included.
- Details of the passwords needed to access the dark archive and the financial and grant tracking databases.
- Pre- and post-transfer inventories. Below I included a small section of both.

Pre-transfer inventory

Directorio de Z:\DATOS\WINWORD\MRP\CHOY

```
06/04/2002  10:41 a.m.    <DIR>          .
06/04/2002  10:41 a.m.    <DIR>          ..
26/01/1999  04:00 p.m.            87.040 ARCIERI.DOC
15/04/1998  12:50 p.m.            13.824 CANAVESE.DOC
10/03/1998  12:07 p.m.            19.456 ENRIC.DOC
11/02/1998  09:30 a.m.            14.848 FPRINI.DOC
26/01/1999  10:26 a.m.            39.443 IRINA.DOC
23/02/1998  02:49 p.m.            35.328 LAZEAR.DOC
18/02/2000  08:53 a.m.           115.712 LIZAR1.DOC
16/12/2003  11:39 a.m.    <DIR>          OLD97
16/03/1998  05:33 p.m.            13.824 RESIDUOS.DOC
20/02/1998  03:05 p.m.            13.312 ZOOLOG.DOC

               9 archivos        352.787 bytes
```

Post-transfer inventory

Directorio de E:\ArchivoAleph\DATOS\WINWORD\MRP\CHOY

```
05/01/2007  02:46 p.m.    <DIR>          .
05/01/2007  02:46 p.m.    <DIR>          ..
26/01/1999  04:00 p.m.            87.040 ARCIERI.DOC
15/04/1998  12:50 p.m.            13.824 CANAVESE.DOC
10/03/1998  12:07 p.m.            19.456 ENRIC.DOC
11/02/1998  09:30 a.m.            14.848 FPRINI.DOC
26/01/1999  10:26 a.m.            39.443 IRINA.DOC
23/02/1998  02:49 p.m.            35.328 LAZEAR.DOC
18/02/2000  08:53 a.m.           115.712 LIZAR1.DOC
05/01/2007  02:46 p.m.    <DIR>          OLD97
16/03/1998  05:33 p.m.            13.824 RESIDUOS.DOC
20/02/1998  03:05 p.m.            13.312 ZOOLOG.DOC
               9 archivos        352.787 bytes
```

Note that the differences between these inventories are the directory dates. The pre-transfer inventory reflects the date in which the directories were last modified and the post-transfer one the date in which they were transferred to the dark archive.

- A document generated by Tripwire including the MD5 hash information for each file in the natural archive's directory. Below I include an example of a file report.

Object:E:\ArchivoAleph\ArchivosServidor\datosaleph\JXM\ANTMISC\OLD\RENOVAR.DOC

Property	Value
Object Type	File
SHA	34057d132407f215a058871e77ae2ec1803f1665
HAVAL	d0bb6fa28005274228afc684fa1d28e0
MD5	499b4dcb9f2ce7a4e1e7604f38718846
CRC32	f16cdc3d
Stream SHA	n/a
Stream HAVAL	n/a
Stream MD5	n/a
Stream CRC32	n/a

One copy of the documentation was given to the foundation's authorities; another one was saved in the documentation directory in the dark archive, and a third one was sent to me by email as a zip archive. Upon receiving and reviewing the documentation I could detect mistakes and oversights.

Oversights

Only two backup read only DVDs were made from the contents of the dark archive: one was given to the foundation authorities, and the second one was sent to me through express mail. Unfortunately the entire contents of the dark archive did not fit on one DVD, so the directory structure was split in two. Therefore, if anything happens to the server and the data is lost, the structure of the directory tree will have to be recovered from the inventories. Another mistake on my part was that I did not ask for backup copies from the networked server, only from the dark archive.

Different from what I indicated in the Transfer and Maintenance Protocol, the only date recorded in the pre and post transfer inventories was last modified date. This is the default date that shows when an inventory command is typed and a list of directories, files, and correspondent dates get automatically generated. To include other dates the command line has to contain specific orders. Lastly, the inconsistency between the directories' dates in the networked server and the dark archive as a consequence of the transfer elicited reflections about the relationships between the nature of electronic files and their file systems, the electronic records' authenticity, and the digital preservation strategy that are discussed in the section *Afterthoughts*. These were yet other events that occurred to the natural archive that I had to document and learn from. In June of 2007 I went to Buenos Aires to complete the dark archive's setup.

ADJUSTMENTS

In mid 2007 the foundation had not been officially closed by the Inspectorate of Justice. Both the networked server and the dark archive were at the foundation's last president's home. The dark archive was running and connected to a computer screen, and nobody had used it since the transfer. The networked server was unplugged, and nobody had used it either. I visited the president's house three times, from June through August, to work with the IT technician and with the president's secretary. We clarified the instructions for accessing the records, made sure that Tripwire's configuration was set up correctly, and tested that the databases were functional and accessible. We also made sure that the passwords to the server and to the databases worked.[230] Shortcuts to the databases and to the shared directory were created to facilitate access.

After reviewing Tripwire's manual, the configuration was corrected to comply with the specifications of the Transfer and Maintenance Protocol. Also, the commands to make Tripwire run were rehearsed and recorded by the president's secretary, who will eventually run the audit and control software during the dark archive's annual maintenance routine. For this, we ran a control test to understand how Tripwire generates results. A copy of the negative report, showing that no changes had been made to any file in the dark archive since transfer, and one of a positive report, consequence of us inserting a comma in a test file, are included in Figure 23 below. Note the box in which severity is 0, meaning that no files had been changed, and the box in which severity is 1, indicating that a file was changed. The positive report shows the details of the modification.

Negative report

	Generated By	Administrador
	Created On	Thu, 02 Aug 2007 14:48:24 -0300
	DB Updated	Thu, 02 Aug 2007 14:38:11 -0300
	Host Name	ARCHIVODIGITAL
	IP Address	127.0.0.1
	Host ID	S-1-5-21-1045716039-1281601870-1452774354
	Policy File	C:\Archivos de programa\Tripwire\TFS\policy\tw.pol
	Config File	C:\ARCHIV~1\Tripwire\TFS\bin\tw.cfg
	DB File	C:\Archivos de programa\Tripwire\TFS\db\database.twd
	Report File	Memory Mapped File
	Command Line	tripwire --check --report-format html --output-file report.html
	Print Command	tripwire --check --report-format html --output-file report.html

	Max Severity	0
	Total Added	0
	Total Removed	0
	Total Modified	0
	High Severity	0
	Medium Severity	0
	Low Severity	0

Positive report

	Generated By	Administrador
	Created On	Thu, 02 Aug 2007 15:04:01 -0300
	DB Updated	Thu, 02 Aug 2007 14:38:11 -0300
	Host Name	ARCHIVODIGITAL
	IP Address	127.0.0.1
	Host ID	S-1-5-21-1045716039-1281601870-1452774354
	Policy File	C:\Archivos de programa\Tripwire\TFS\policy\tw.pol
	Config File	C:\ARCHIV~1\Tripwire\TFS\bin\tw.cfg
	DB File	C:\Archivos de programa\Tripwire\TFS\db\database.twd
	Report File	Memory Mapped File
	Command Line	tripwire --check --report-format html --output-file report.html
	Print Command	tripwire --check --report-format html --output-file report.html

	Max Severity	100
	Total Added	0
	Total Removed	0
	Total Modified	1

Modified Object: E:\ArchivoAleph\backup.log

	Property	Expected	Observed
	Object Type	File	File
(*)	Size	3.600	3.601
	MS-DOS Name	backup.log	backup.log
(*)	SHA	ed6b67d4f17e93ba0ffe49a67577a79b7ab026dd	59447808fd7377232501cf2cd769da5c09f209fe
(*)	Write Time	Viernes, 02 de Noviembre de 2001 08:17:16 a.	Jueves, 02 de Agosto de 2007 03:03:42 p.

Figure 23: Negative and positive report summaries from Tripwire.

PS

In March of 2008 it was decided that the contents of the networked server were going to be erased and the server donated. The dark archive will remain in the same facility in which the paper records are stored, inside a small air conditioned vault for electronic media storage. The networked server was cleaned up on April 5, 2008 by the same IT technician who set up the dark archive. I sent him instructions to follow and document pre-and post- cleaning procedures, among which I included the request to create a backup DVD from the networked server and a final inventory to register what the networked server held at the moment of disposition. In this case, I specifically asked that the inventory include last modified, last accessed, and files' creation dates. The inventory and the documentation, including a step by step account of how the networked server was cleaned are now stored in the dark archive, I have a copy as well.

MONITORING THE DARK ARCHIVE

The annual monitoring routine was planned to check the integrity and authenticity of the dark archive's contents throughout the records retention period. It was also suggested to verify the server's performance. For the latter, it was planned that post transfer MD5 hashes would be compared to new ones on an annual basis to determine if file modifications had occurred, in which case the server access and file change logs will pinpoint who made the changes and when.[231] It has been foreseen that Aleph's records could be accessed during the records retention period, but that they should not be modified. If access to the files or the databases occurs, this will be recorded by Tripwire, and if modifications occur, a report like the one shown in Figure 23 will be issued when the audit checks take place. If needed, original versions of modified files could be restored from the backup DVDs. These audit checks to detect modifications were to be conducted by the last president's secretary on an annual basis. Now that the dark archive is in a secure storage facility the integrity and authenticity controls are less pressing, but they will have to be conducted if and when the dark archive is accessed for consultation.

I spoke with Shane Williams, the UT Austin School of Information System's Administrator, about whether maintaining a server (as hardware) shut down for the next ten years is a good idea.[232] Of concern is that in ten years the hook-ups and cables from the server to the screens and keyboards may change. This will be solved by incorporating that concern as a technology follow-up, just as I will follow-up the behavior of the files and the DVD copies. Regarding how the server's mechanics will respond, it is not possible to predict what will happen. A positive aspect is that it will be kept in a clean and air-conditioned environment.

VIRTUAL MIGRATION

I have stated that the dark archive cannot remain as a black box. As new hardware and software becomes available, future upgrades of the dark archive will be evaluated in relation to the need to move the records and databases forward. This in turn will depend on final decisions about the destiny of the archive beyond the ten year retention period. These evaluations remain under my responsibility and will be conducted as frequently as new software for Windows—OS and Office versions—is released. I will conduct these tests virtually. From the back up DVD in my custody, the same sample of records used for the file rendering test and the databases will be used to conduct file rendering and functionality tests. Changes will be recorded and the information will be used to make migration decisions.

The first file rendering test was conducted in February of 2008 with Windows Vista and Word 2007. The only change noted with respect to Windows XP/Word 2003 was that two character codes do not show as glitches in the new version. A column with the results was added to the column in the File Rendering Testing Table included in Appendix V. The virtual migration strategy is a cost effective complement to migration on demand. It means that when a decision about migration is made, we will have the right information about what files need to be migrated and to what format. I will also be in charge of conducting annual evaluations of the backup DVD media. This entails being aware of new storage technologies, determining whether the media needs to be refreshed and to which new format, and then conducting the copy.

Afterthoughts

After the transfer was completed I had time to reflect about this experience and what could have been done better. In this section I focus on file and directory properties, how they change as consequence of use and during transfer to the dark archive.

FILE PROPERTIES AND AUTHENTICITY

File properties are metadata about characteristics, use, and location of files such as: file size, creation date, last modified date, last accessed date, file path, number of words, author, edit time, etc. (the number and type of properties varies depending on the creating software). All of these elements may used to determine the authenticity of an electronic record. Moreover, file properties are considered legal evidence, and current litigation recommendations insist that they should be preserved unchanged once an electronic document is produced.[233] In this research I have used file properties, both observed and extracted as metadata, in the formation process study and to build the Metadata Timeline to indicate when different technologies started being used. To build the yearly sets for the appraisal method, I used the last modified dates embedded in the files as references.[234] In the Minimalist Preservation Strategy I indicated that the archive had to be kept "as is" for authenticity and integrity purposes. Therefore, the way in which file properties behave in uncontrolled record-keeping environment, the consequences of transfer and transitions on them, and the way in which they can be managed once the records are archived needs to be comprehended.

File properties are used to manage files both by the file system in the local operating system and by file management software. This metadata is embedded in the

files per se and recorded in the file system. Some file properties depend on the configuration of the operating system in which the files are created and used, others on the configuration of the software that creates the files. Depending on the file format, the software with which a file is created and opened, and the particularities of the different file systems, file properties are recorded and displayed differently. For example, a .doc file opened with Open Office displays file properties differently than if opened with Microsoft Word, and MS-DOS files record file properties differently than Microsoft Office Word files. Moreover, the way in which the system records file properties may vary from version to version of the same operating system.[235] Depending on the actions done with the files, whether they are copied or saved to and from a directory or to and from another storage system such as a server or a portable hard-drive, file properties may or may not be overwritten.

In the following explanations I point specifically to Microsoft Office files because that is the environment in which the natural archive was contained during the last ten years and it is the one in which it will be stored for the next ten. The explanations are based on my observations of how file properties change as consequence of different actions such as copy, save, save as, and transfer to other directories, to a different hard-drive, or to a server with a different operating system. The next figure illustrates how file properties in Word 2003 documents are displayed from the General and the Statistics tabs in the File properties menu.

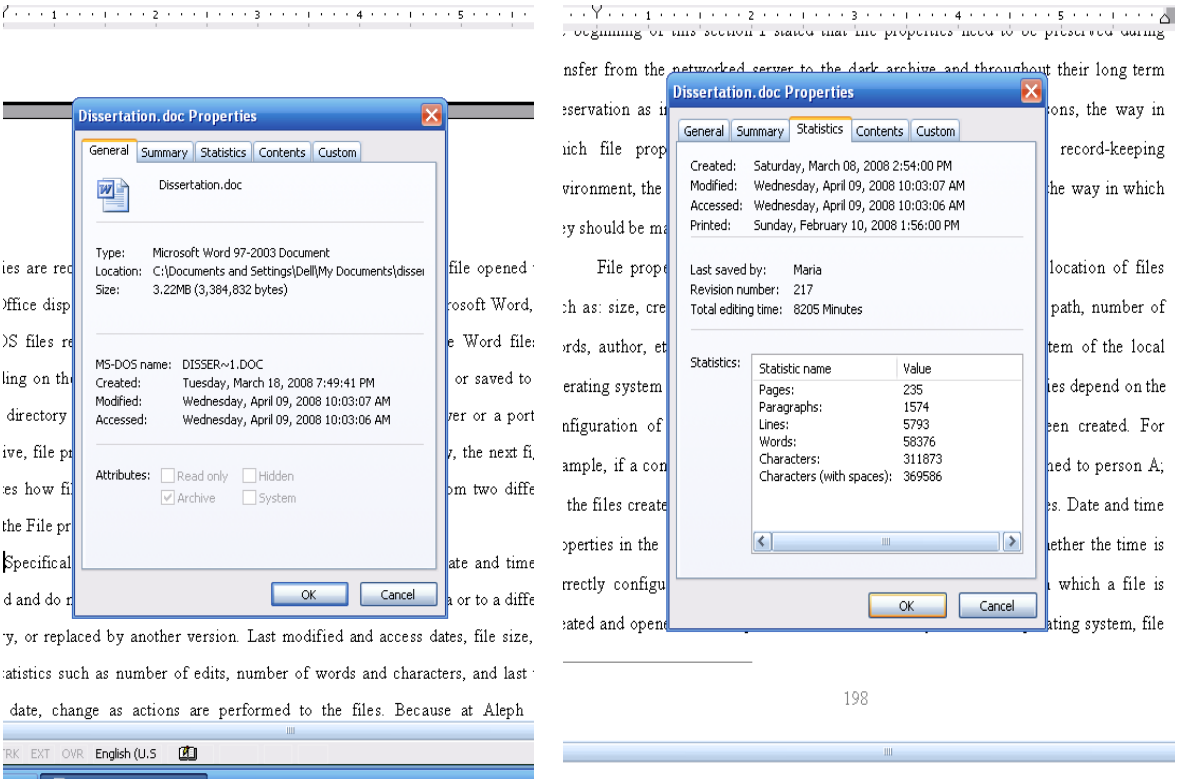


Figure 23: Display of file properties in a file created in Word 2003 for Windows XP.

Note that the created date in both views is different; the one displayed in the General tab shows the date in which the file was saved to its current location, and the one in the Statistics tab the date in which it was either created, or saved to the location in which it was before it was saved to its current location. When Microsoft Office files are managed within a Windows system, the file properties display is a hybrid of both the metadata embedded in the file and that one in the file system.

When a file is created the date and time that comes from the local operating system is recorded in the file and does not change unless the file is saved to a different directory or transferred to a different hard-drive. Last modified and access dates, file size, and other statistics such as number of edits, number of words and characters, and last

time printed date, change as actions are performed to the files. Because at Aleph staff members worked between their local hard-drives and the networked server and files were shared, saved, replaced, and transferred across these systems, the creation date may show a time-stamp later than the last modified date. Another example of how date and time stamps can be unreliable is that they are recorded according to the local time zone and may not be precisely configured in the computer's operating system.

In terms of authorship, most computer-stations at Aleph were assigned and the name of the assignee is the one recorded in the file properties. However, if a document was copied by another employee to his or her sub-directory for modifications, the authorship field in the file properties will indicate the first author. If instead the document (through the copy and paste function) was used as a template and saved under a new file name by another staff member, his or her name will display as author. However that could change if the creating program (Word, WordPerfect, etc.) was configured with a different name in which case the files will have that name in their properties.

Directories also have creation, last modified, and last access properties. As far as I have been able to experiment in the Windows operating system, when directories are copied/transferred to other storage media (such as a DVD, a hard drive or a server), their time stamps are overwritten to the date of the transfer. Also, every time a directory is opened to include, remove, or save files within it, the directory's last modified date changes. The last access date for the directory changes every time a file in the directory is opened. Over the years the directories of Aleph's natural archive were copied to or transferred from other storage media. Therefore, on the networked server there is no way to assert the date in which a directory was created; one can only infer that it was circa the earliest file modified date within the directory.

When the files from the networked server were transferred to the dark archive, the file system in the dark archive recorded the date of the transfer as the files' creation and last accessed dates, which is different from the metadata embedded in the files themselves and remained stable during the transfer process. When the directories from the networked server were transferred to the dark archive—Windows OS—their time stamp changed one more time; these changes are documented in the pre- and post-transfer inventories (see section *Transfer*). Some days later a backup of the contents of the dark archive was made on DVDs and the dates in the directories were once again overwritten. In case of having to restore the files from the DVD, these changes will have to be documented. Most importantly, for future transitions, other file transfer and backup systems such as magnetic media or disk image copy that assures the preservation of the file properties will have to be implemented.

In post-custodial mode, the records and databases in the natural archive will remain in the hands of their creator through the next ten years, during which it is anticipated that they may be accessed in case of special need. Even if not modified, if the records are accessed, saved after being accessed or transferred back and forth, some file properties will change. The audit and control software will capture access and changes in location, and the pre-and post- transfer inventories will attest to the location of the files and their last modified date as they were when the institution closed. In this case, the redundancy of the documentation requirements in the Transfer and Maintenance Protocol covers for mistakes and omissions.

As a consequence of learning from this experience, other options to secure that no further changes occur in the file properties in the natural archive were considered. One was to make the entire dark archive read only. However, applications will not function

well with read only files, and this archive may still need to be accessed. Another alternative, to preserve and freeze the natural archive as it was before the transfer, was presented to me by Shane Williams and by Dr. Patricia Galloway. It involves creating a bit-by-bit disk image copy of the contents of the networked server. This is a technique used to clone and backup hard-drives. It can be used to restore a hard-drive's structure and contents but is not useful to access the files.

I analyzed this alternative from legal and practical perspectives. Since during the retention period the archive might have to be consulted, the natural archive needs to remain functional. If there is any legal occurrence and a judge issues a subpoena to access the records or databases, the functional archive will be the legally bound archive and not the disk image copy. Freezing the archive when it still may need access entails creating copies in some way or the other; something that would complicate the archive's management and legal reliability. I judge that such image would have been very useful vis-à-vis the split of the directory tree in the backup DVDs, and as pointed out by Dr. Galloway, may still be made to serve as a record to follow the changes that may occur in the future. For the moment, the changes that might occur as a consequence of processes will be documented. Despite their unstable nature, and the fact that their changes are extremely confusing to follow through, file properties—interpreted in the context of the characteristics of the archive at hand—can provide clues to the archive's authenticity. And still, the characteristics of the natural archive highlight the blurry boundaries that exist between electronic records as active systems and electronic records as archives.

Creation, Use, and Preservation of Electronic Records

Upon completing the preservation strategy, practical lessons and theoretical implications are discussed.

Marks in the archive

Implementing the minimalist preservation strategy helped me better discern the relationships between records creation, use, and preservation. In the context of undisciplined record-keeping practices, from the time of creation each use leaves whatever mark in the record's metadata the creating software allows, which in turn may erase previous metadata. During the normal course of affairs, as records are copied and transferred from one hard-drive or storage medium to the next, these changes are not an issue, they are not noticed by the user because records are tools and paths to outcomes. During the transfer of the records to the dark archive and the production of backup copies, the changes in properties highlighted how archiving and preservation decisions may also leave marks on the digital objects. At the archives threshold, the question is how to address the consequences of transitions both from practical and theoretical points of view.

Unlike paper, electronic records are dynamic and malleable entities; and unlike electronic records created and maintained in disciplined record-keeping environments, records maintained "naturally" may change their properties to the point that their creation date—unless explicitly recorded as content—may become uncertain. From here on; just as I decided to preserve the structure and the wholeness of the natural archive, I made the decision to stabilize the archive and document each activity that may occur. This decision did not come without dilemma. I considered whether to stop further imprints on the

archive through actions of preservation and access; or to let these changes occur and documenting them elsewhere to provide an audit trail. Because it is in the nature of digital files to register and overwrite changes, I clarified the meaning of file properties and made an effort to make them explicit as time goes by. In this strategy I understand that the archive's formation process does not stop. At least until the record retention period expires, I am preserving both content and processes.[236] . Below are the strategies that will be followed if the decision to give public access to the records beyond the records retention period is made.

Migration on demand and virtual migration

The minimalist preservation strategy coincides with the strategies outlined by David Holsworth in the installment "Preservation Strategies for Digital Libraries" of the Digital Curation Manual, which reflects lessons learned from the CEDARS project.[237] This project recommends upgrading the technical environment in which digital objects are maintained and to migrate them "on demand" when there is a need. This differs from the strategy of migrating records each time a new version, a major revision, or a discontinuation of the creating software occurs.

The key to migration on demand is to maintain the original bitstreams and their representation information over time, which will allow identifying an adequate migration protocol for these objects in the future. In the minimalist strategy bitstream preservation is assured because the records are stored "as is" in the dark archive. The representation information of a digital object can be as simple as the object's file format or as complex as the file structure and position of every digital element such as the case of a web-page.[238] In the case of Aleph's natural archive, the representation information is recorded as groups of similar record types in the Metadata Timeline and individually for every file

in the pre- and post-transfer inventories. In the case of the databases, the representation information is understood as documentation about the object's dependencies—software version and compatible OS—and its development code. The former is recorded in the Metadata Timeline and the latter is kept in a designated directory in the dark archive.

I call *virtual migration* my contribution to the migration on demand strategy. By virtue of using a copy of the digital archive (or a sample of representative digital objects) to perform rendering and functionality tests in up-graded technical environments every time that a new technology emerges, the hibernating archive is being moved forward virtually. These tests are complemented with migration tests and include information about the migration software available and the leap that it implies. If and when the need to update the technical environment or to migrate the records and systems arises, the information about what needs to be migrated and how will be available. In any case, the migration will be completed by using copies of the files, and the original bitstreams will remain as they are.

Continuum

While future representations of the records through migration or due to other digital preservation and access methods will have their own authentication and iterations (and some of them will happen at different times for different file types), the natural archive preserved "as is," or its disk-image, will provide the evidence of what the natural archive was.[239] The changes undergone by the electronic records before being deposited in the dark archive, and the constant preoccupation that maintaining and transforming them will imply in the future, indicate that the natural archive and every other electronic archive are in a continuum with their former states.[240]

The preservation process highlights the fact that post-custodial scenarios are based on mutual trust between the creator and the archivist. Aleph's archiving process shows that the organization trusted me as an archivist to lead a process that protects the archive and them as legally bounded institutions. In turn, their will to comply with my specifications during the retention period shows that they will protect the authenticity and integrity of their archive. While so far this is a successful post-custodial story, it poses the question of whether others will be able to afford, even with a minimalist preservation strategy, supporting their own archives over time.

Inevitably, I compared the results of the minimalist strategy with the option of loading the records in an institutional repository system (IR) that I considered at the beginning of the archiving project. Certainly, the representation of the archive would have been very different. In the IR the staff members would not have loaded personal records or fragments for which they could not add precise metadata and all the applications and remnants of digital objects would have not been preserved. There would have been a weeding process and the representation of the archive would have shown a domesticate archive.[241] On the other hand, risks of changes in file properties during the records retention period would have been minimized. In hindsight, if I would have proceeded with my first plan, this dissertation would have been a different one and the archive would have turned into a very different—less complex one. Now I know what the archive is; I would not have known otherwise. Besides making me fully aware of my intervention in the natural archive's formation process, the preservation experience allowed me to foresee how future preservation measures and transitions are already shaped by what I have done.

PRESERVATION RESEARCH FOLLOW-UP

To monitor the long term preservation of the digital files in the dark archive, I will continue the horizontal time study of the behavior of different file formats and databases. Tracking the point at which these files and systems become un-renderable or show formatting and character display changes due to updates of their native software or operating system will establish when and if they should be migrated. More detailed research has to follow to document the correspondence between the different glitches and the character codes. In this dissertation I have focused on MS-DOS Word and Word for Windows files; the natural archive has other file formats and applications whose rendering and prospective migration time needs to be documented. Among those are image files, Lotus 1-2-3 and Excel files, .ppt, and .pdf, different types of databases, and some DOS-based applications. I will complement the study by testing migration tools and documenting changes consequential to migration to progressively build up this research towards creating a migration decision making tool. I will also continue researching file properties' characteristics and changes as well as tools for analyzing, recording, viewing, harvesting and preserving that metadata. These studies are categorized as digital forensics.

PART V: ALEPH IN THE ARCHIVE

A Natural Electronic Archive

Creation of "natural electronic archives" involves a set of *ad-hoc* practices developed as people adjust to, learn, and use information technologies. Natural archives can be created by one person or by a group of people with different levels of technological understanding and using different tools. In a natural archive each record creator decides on naming conventions and organization for files and folders, spontaneously and, sometimes even consistently; in other words, each individual applies mnemonic rules or decides on the spur of the moment. In their origin, natural electronic archives are not meant as archives; they are individual expressions of records' organization and of personal and/or work practices in which records are tools, templates, and final versions.

Users create, re-invent, and leave behind natural archives or parts of them in iterative fashion. Through these iterations, archives evolve towards more structured forms or become interrupted, and during these processes, pieces of the archive are left on storage devices like discarded artifacts on an archaeological site because there is no time or need to go back to old files that cannot be found or applications that cannot be opened. This constant advance and backtracking in the archive is intensified by emergent technologies and ways of creating and storing records, and the appearance of new users who enter the learning circle. A natural archive is an accretion of trials and errors, the evidence of which resides in the directories, files, and applications left in the storage space and in the lack of consistency in naming, organization, versioning, keeping, and discarding that characterizes them.

What motivates disposal or retention is each person's decision about whether his or her records become useless, could be relevant for future work, or are fundamental to them what motivates disposal; which for the most parts happens at the time in which the process that is occupying the record's creator takes place or ends. But in disposition decisions there is also a component of carelessness, of letting it go, because people forget the records that they have or can't find them, and in many instances disposition happens by mistake. At Aleph, most people hesitated to look into other people's records and did not dare to destroy each other's records. All of this speaks of the resilience of electronic records in the natural archive, which survived even beyond the intent of their creators.[242] From a digital preservation perspective, natural archives are also resilient. In my case study it is possible to render today. Albeit with some glitches in the case of MS-DOS files, records that had gone through several hardware updates, two major changes of operating systems, and at least five creating software versions.

At this point I suspect that natural electronic archives are a consequence of hybrid environments; of the existence of a formal structure—paper files or databases—out there within which the finished record exists as a valid entity (in Aleph's case it would be the project file). In a hybrid records environment, the same creator that maintains a perfectly organized paper archive may or not maintain a perfectly organized or complete electronic archive.[243] Either on individual hard-drives or in a shared directory, records in the natural archive constitute personal collections not explicitly prepared to be accessed by others. Moreover, records in the natural electronic archives are many times difficult to retrieve by their own creators, inspiring frustration and abandonment of the existing logic of the archive which might be re-started. The point is that natural electronic archives are in continuous transformation.

As they evolve, natural archives are not managed according to strict retention schedules or legal impositions; [244] their records are not identified as having other than the work and personal dimensions intended by their creators. In their genesis and throughout their development, they are oblivious to a priori archival models of the records life cycle or records continuum.[245] However, these models are useful to compare and contrast against the characteristics of the natural archive. The records life cycle model marks strict phases which in practice are formalized as records retention periods. After records become archives their functions change as they no longer fulfill the role for which they were created. In this conception, management of records is linear and efficient. In the records continuum model transitions towards other dimensions are less strict, simultaneous, and are dependent on the context in which records are used and on how they are perceived by individuals and society. Records have diverse meanings, they can traverse through different dimensions at the same time, and they are managed within that complexity.

Witnessing Aleph's records transition to become archives, I did not see a rupture such as the one proposed by the records lifecycle construct.[246] Instead, the natural archive drags with it some of the same fundamental concerns about privacy, confidentiality, administrative transparency, and emotions that shaped the institution while active. In this sense, the evolution of the natural archive better matches the passage of records towards the different dimensions of the records continuum model. In a country like Argentina, where the concept of records life cycle is not familiar, it is hard to impose on the creators the notion that as the institution closed, its records don't have the same function. Furthermore, the digital appraisal method showed that in a natural archive records' roles are multiple and contextual.

In an environment without explicit recordkeeping rules, bits and pieces of text are ubiquitous inhabitants. Either shared by different members of a network or used repeatedly by their creators, they constitute the core of many records. The repetition of fragments, afforded by the cut and paste function of the text editor, speaks as much of provenance, group collaboration, and fair use, as of hierarchies and organizational culture. The presence of these pieces highlight processes and steps, and in a shared environment the roles played by the different individuals in those processes. Mixed, intertwined, combined; the records and the myriad of digital objects that may surround them are both context and content as they explain each other's origin and uses.

Natural archives are ubiquitous: present in personal and organizational computing storage devices. And yet, they are only acknowledged after the fact, when there is a need to find a file created a long time ago, or a decision has to be made about their fate.[247] In the natural archive everybody is represented, from the receptionist to the president and the part-time employee. Within these personal collections, the records reveal the multiple dimensions of their creators: personal, professional, academic, civic, and spiritual.

Discussing memory, technologies of inscription, archiving, and psychology, Jacques Derrida considers the archive as "impressions" that may in the future constitute the concept of archive.[248] This conception maps the one of the natural archive which, formed by scattered texts (as impressions) left on the networked server, has the potential of representing the people and the organization that created it. In this case, the aggregation of repetitions and fragments of texts, complemented and informed by the vestiges of digital materials, allow viewing today the *Aleph* of the past—both as an organization and as individuals—"on the fly,"[249] "without super-positions and without

transparencies." As for the future, preserving the natural archive as it is will ensure its use and the continuous representation and interpretation of *Aleph.*

CONTRIBUTIONS

As a final synthesis I comment on what I consider the main contributions or focal points of this research. Taking Dr. Patricia Galloway's advice, I use the reactions and opinions of colleagues about this work as points of reference to discuss my findings.

A consistent question that I was asked was how I resolved being a researcher and an object of analysis: a staff member and the archivist. Such a research approach can be framed as participatory action research, or action research, except that I did not go into the field with the express idea of conducting research, but stumbled over the topics.[250] I did not doubt the value of the archive, the question was indirectly posed to me and required research because there were no answers from the archives field for cases such as Aleph's. Beyond that point, I explain my steps extensively in the first part of the dissertation and clarify from the start that my intention was and is to preserve Aleph's archive. Archives, especially private ones, persist because of someone's intent, and I should add that coming from the field of preservation and conservation, that desire is a main element in our research motivation. Ultimately, it is the rigor of my research that will be judged and will show whether my results are sound.

In the first part of this dissertation I tell the story of a digital archive, and by doing so I describe how an organization matured in the use of information technologies. Aleph was a pioneer in the implementation of networked personal computers in the work-place in Argentina and the narrative of the digital archive's formation process provides a unique view of technology adoption in the midst of an adverse economical and political

environment. Argentina is not a developer of IT; it copies and clones, it adopts and adjusts. Aleph's story shows the drive to absorb and implement; the isolation in which IT decisions were made; and the difficulties in thinking through and communicating the possibilities of these technologies. I conclude that these difficulties were a consequence of the novelty of the technology as much as of the gap in comprehension and sense of ownership that exists when users and adopters do not form part of a culture of technology development.

Sections of the story highlight the transition between paper and digital records. In terms of uses applied to electronic records, Aleph's narrative comes full circle; from conceiving electronic records as tools to generate official paper documents to considering them as official communications when the organization decided to grant scholarships through the Internet. And yet, the recognition of electronic records as official records was not completely internalized. This manifests the characteristics of hybrid paper/electronic archives, a phenomenon that has not been sufficiently studied from an archival perspective nor given the importance that it deserves when considering records' values during appraisal. Indeed, it is the hybrid character of this archive that evidences the unity and complement between the paper and the natural electronic portions.

Critical reactions to the concept of natural electronic archives that came from archival professionals helped shape the concept. One critique suggested that the name of "natural archive" was redundant, as all archives are created naturally during the course of affairs.[251] Another pointed out that if natural archives are uncontrolled and most likely incomplete they will not reflect well the organization or person that created them. These two remarks have common grounds.

The definition of archives in the SAA glossary states that, "…archives are characterized by an organic nature, growing out of the process of creating and receiving records in the course of the routine activities of the creator (its provenance)."[252] And yet, this definition does not account for the activities of records managers and archivists and how these shape the archive; for example by removing some or many of the records that derive organically from the work or personal processes.[253] Archives that derive from strict record-keeping systems are efficient and they may be precise in the representation of the organization, but the fact that they are carefully architected removes the simplicity of direct representation. Also, the definition does not account for curated archives, made to represent certain aspects of a community, a person, or a society.[254]While records are natural outputs of transactions, once they are regulated, controlled, and weeded; the archive that results loses the properties that I describe as belonging to a natural electronic archive. The concept of a natural archive highlights the organic nature of archives by pointing to the way in which people create and accumulate records in myriad forms stressing those in which their will is somehow removed from their actions.[255]

In the natural archive, the dilemma of selection belongs to each individual creator.[256] Is this more or less legitimate than to apply a series of standards to selecting what should stay? What type of approach renders a better view, the selection practiced by the creators or the archival selection? During my years as a doctoral student I did proxy research at the Harry Ransom Center for scholars who could not travel to Austin. There I had the opportunity to search through archives in which every single piece of paper counted—envelopes, notes, brochures, lists, versions, repetitions—everything is collated and offered to the user who ultimately decides and interprets. As a mediator between the archive and the researcher, I could observe how all the pieces were important and made a

difference. Researchers did not discuss with me whether or not the archive was complete, they were just thrilled with every piece that I found. Informed by my experiences with literary collections, my approach was not to judge "a priori" and discard or select the electronic records but to find the archive that was hidden on the server. For this, I used archaeological methods because they are designed to deal with vestiges and gaps,[257] and devised a data driven digital appraisal method to extract knowledge from the remaining text records.

Designing a digital appraisal method and, with the help of Dr. Hai Bi, the tool with which to conduct it was a major part of this dissertation and remains an ongoing research topic for me. Electronic records afford such studies, and archivists have much more interesting and engaging opportunities than they did before. My contribution is to open the door for text mining and visualization to be used in archival appraisal, to put forward the idea that electronic archives can be analyzed combining archival and data driven methods,[258] and that because these are archival problems (including those incumbent to the way in which digital records are represented and re-interpreted over time), archivists are the right professionals to create the appropriate tools and conduct the analysis.

My colleagues' observations about the digital appraisal method were very fruitful. For example, questions about the internal validity of the average of cosine similarities encouraged me to work with cosine similarity distribution curves which can also be used to look at the results from different perspectives. Other comments and questions revolved around the scalability of the method and whether it can be practically implemented like other appraisal procedures such as functional analysis or macro-appraisal. About scalability, an IR scholar suggested that over the years the increase of records not related

to the organization or what we denominate non-records, will most likely distort the results. A point to consider however is that in the conception of natural archives non-records are not spam; they are records that people keep over the years.

On the supposition that staff members would have increasingly generated more and more non-records such as, for example, personal ones, the representation may have reflected a less and less focused organization over the years (a consequence of doing private matters in the work-place). Or, given the regulatory trends that have emerged lately, another supposition is that the organization would have implemented measures to prevent staff members from creating and maintaining personal records in the work-place and have enforced a rigid record-keeping structure. In that case the natural archive would have morphed into a domesticated one.[259] There is no telling what a natural archive can become, but preserving it "as is" allows for the exploration.

I designed this appraisal method with a theoretical goal, to determine whether evidence can emerge from unstructured and un-documented records. Using it as a regular appraisal procedure will involve adjustments of the research design and the tool, as at present both elements are geared towards this data set. But beyond this particular application, the experience informed me about the possibility of merging appraisal and organization of unstructured text corpora. One useful observation made by a Corpus Linguistic professor about the project was that the results have an IR component because one could start by studying these records based on those shared between staff members. So looking at the matches between the records of the cultural assistants and the cultural manager will show the pool of projects that the area was involved in at a given time. This method can thus be used as an introduction to explore not only the relationships but the activities that drive them.[260]

Along with the minimalist preservation strategy, I introduce a post-custodial solution for digital archives. Included is the Transfer and Maintenance protocol that can be used as a template to ingest and stabilize records throughout the period of records retention. In this case I want to highlight the value of the archival way of doing things, the lessons learned from the sometimes painstaking but necessary routines that occupy archivists throughout the processes of appraisal, acquisition, and processing. It was the redundancy of the protocol that allowed the documentation of transitions and changes despite my omissions and mistakes. Another colleague was surprised that files from 1993 could still be opened. In the future, the virtual migration strategy will allow following up these files through future platforms and testing the archive's resilience.

All the conclusions and the findings are a consequence of the hands-on approach to looking at digital objects, to analyzing them as cultural material, a practice that I embraced through my years of work in the field of Heritage Preservation. They also derive from the way in which I was trained in Digital Archiving and Preservation Administration in the School of Information at the University of Texas at Austin, where as students we embrace the unknowns and uncertainties related to digital preservation and archiving and are encouraged to find answers based on observations and experiences. Throughout my presentations of this research I received all sorts of comments, from "this is average work" to "it is genius" and everything in between. The comment that I valued most of all was "your dissertation was a lot of work." I find that comment the most accurate and complementing. It was a lot of work and it will be a lot of work to proceed with what is left to explore.

The one question that has not been brought up yet is: Are you keeping it all including applications, email, databases, non-functional and functional applications,

scripts, etc.? YES, all the contents of the natural archive were necessary to tell the story of the archive and the organization and to establish the records' authenticity. And yet, despite my involvement, this research, and my insistence, it is not in my power to decide whether or not to preserve the natural archive for the long term. I do hope that this dissertation indicates that this archive is vital to maintaining Aleph's memory so that in the future it does not become a myth. Beyond this, I deeply appreciate the incredible opportunity that I was given; to have a real archive to work with. If nothing else, by allowing me to work with their archive foundation Aleph has once again contributed to the advancement of knowledge.

Austin, Texas, April of 2008.

Appendix I: Interview Protocol

Questions for staff members

- When did you start working in the organization and up to what date? What was your function? Whom did you report to and who supervised you? Who did you work with within the organization? Tell me about the changes in your job.
- What type of records did you create regularly? How did you name them? Where did you store them? Did you have any naming convention or consistent record-keeping practice?
- What was your records creation routine? Did you solely create records individually, collaborate with other people, or edited/supervised records produced by others?
- Were there institutional record-keeping and record-making policies? For example, how to keep or discard records? If so, who established them? Was it different from the paper archive?
- What type of information technologies did you operate for your daily work?
- Did you know about systems administration and backup policies?
- Did you take part in decisions regarding the selection of systems or the frequency of technological updates?
- What were your criteria to keep or delete records? Specify according to different types of records that you created on a regular basis: email, web-site pages, administrative correspondence, reports, memos, appropriation requests, meeting minutes, etc. Also according to whether they are paper or electronic.

Questions for systems administrator and IT consultants

- When did you start working for the organization? For how long?
- What project did you developed? Who intervened in the development?
- What was the language/platform/software used to develop these systems?
- Indicate considerations when these decisions where made
- Did you leave any user manual? Tell me about the process of staff training
- How was the computer network set up? Detail its historical evolution and technical components.
- Who had authorizations to see, edit, and delete which records?
- What was the network's backup system like?
- Was there any type of policy for record-keeping or record-making?
- Is there an inventory or list of applications used over time?
- Are there records where changes, purchases or licenses have been recorded?
- If any, describe frequency of renovation for hardware and software.

Appendix II: Metadata Timeline

Year 1987-88					
Active records	**Name**	**Version**	**Version Date**	**Technology discontinued/or upgraded**	**Source**
	System				
	Hardware				
	AT (clon?)	286	1982	1986	Interview
	Network OS				
	Novel Netware	286	1986	1989	Interview
	System's programming language				
	dBASE	3	1984	1988	Interview
	PC				
	Hardware				
	IBM PC				Interview
	PC OS				
	MS-DOS				Interview
	Writing software				
	Word Star				Interview
	Word Perfect for DOS				Interview
	Word for DOS	3	1986	1987	Interview/found diskette from program
Year 1989/1990					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	PC				
	Spreadsheet software				
	Lotus 1, 2, 3	2	1985	1989	server

Year 1991					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	Network OS				
	Novell Netware	386	1989	1993	Interview
	System's programming language				
	Clipper	87	1987	1990	Interview
	dBASE	3 or 4			Interview/server
	PC				
	Hardware				
	AT	286	1982	1986	accounting book
	AT	386	1986	1989	accounting book
	OS				
	MS-DOS				server
	Writing software				
	Professional Write				diskettes from tare
	MS Word for DOS	5	1989	1991	server
	Spreadsheet				
	Lotus 1 2 3	2	1985	1989	server
Year 1992					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	PC				
	Hardware				
	AT	286	1982	1986	accounting book
	AT	386	1986	1989	accounting book
	Notebook Toshiba TI1800/60				accounting book
	Writing software				
	MS Word for DOS	5	1991	1993	server
	Spreadsheet				
	Lotus 1,2,3	2	1985	1989	server
	Quatro Pro				server

Year 1993					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	PC				
	Writing software				
	MS Word for DOS	5.5	1991	1993	server
	MS Word for DOS	5	1989	1992	server
Year 1994					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	Hardware				
	AST Premium MTE 423				accounting book
	OS				
	Novell Netware	3.11	1989		accounting book
	PC				
	OS				
	Windows	3.0 or 3.2	1990 or 1992		inference
	Writing software				
	MS Word for DOS	5.5	1991	1993	server
	MS Word for Windows	6	1993	1995	server
Year 1995					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	PC				
	Hardware				
	VTC (clone?)	486	1989	2007	accounting book
	Writing software				
	MS Word for DOS	5.5	1991	1993	server
	MS Word for Windows	6	1993	1995	server

Year 1996					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	Hardware				
	ACER Altos 900				accounting book
	PC				
	Hardware				
	Upgrade to Pentium		1993	1999	accounting book
	OS				
	Windows 95		1995	1998	found installation disks
	Writing software				
	MS Word for DOS	5.5	1991	1993	server
	MS Word for Windows	6	1993	1995	server
	MS Word Office 95	7	1995	1997	server
	Spreadsheet				
	Lotus 1,2,3	2	1985	1989	server
Year 1997					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	System's programming language				
	Clarion	2			interviews/server
	Internet				
	Hardware				
	SUN				accounting book
	PC				
	Email client				
	Eudora Pro	4.1			
	Writing software				
	MS Word for Windows	6	1993	1995	server
	MS Word-Office 95	7	1995	1997	server
	Spreadsheet				
	Excel Spreadsheet	7	1995	1997	server

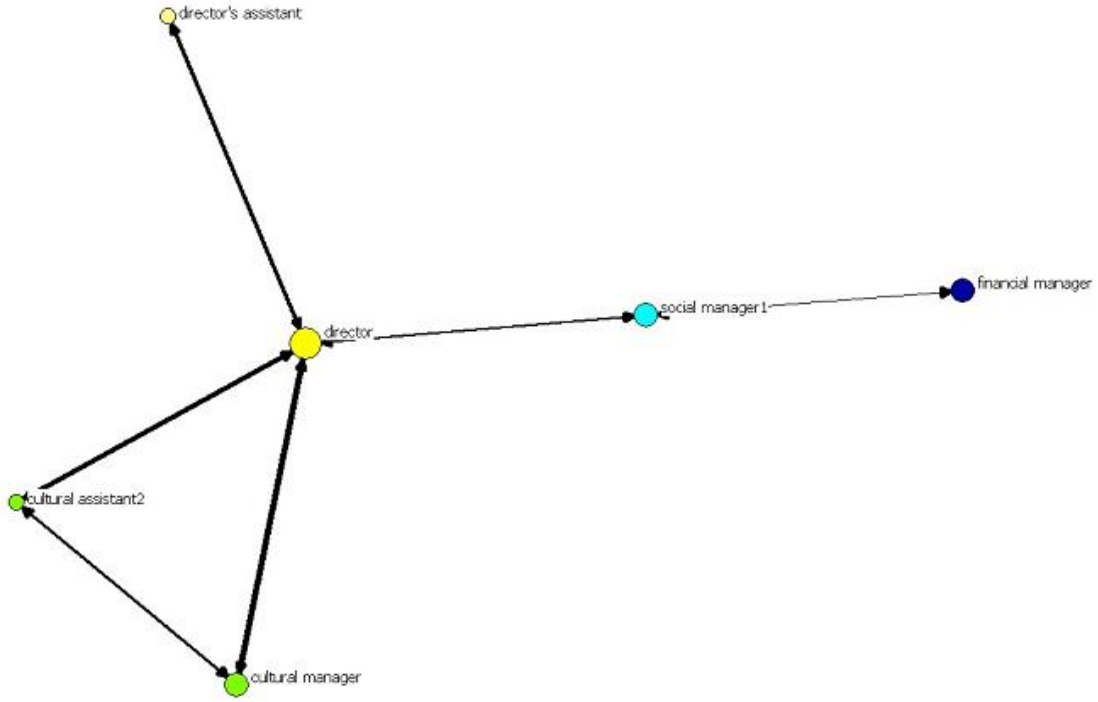
Year 1998					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	OS				
	Windows NT server	4	1996		interview
	PC				
	Hardware				
	Notebook AST M5260				accounting book
	Notebook Compaq Pentium	II	1997	1999	accounting book
Year 1999					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	PC				
	Hardware				
	Pentium	II	1997	1999	accounting book
	Pentium	III	1999	2003	accounting book
	OS				
	Windows 98		1998		found installation disks
	Writing software				
	MS Word-Office 97	8	1998	1999	server
Year 2000					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	Hardware				
	HP Server LC3	Pentium III			accounting book
	OS				
	Windows NT server	4	1996	2000/2001	accounting book

Year 2001					
	Name	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	Hardware				
	HP Netserver	Pentium III			accounting book
	OS				
	Windows 2000		2000	2003	found installation disks
	PC				
	Hardware				
	Pentium	III	1999	2003	accounting book
	Pentium	IV	2000	2008	accounting book
Year 2002 - 2006					
Dark archive	**Name**	**Version**	**Version Date**	**Technology discontinued**	**Source**
	System				
	Hardware				
	HP				Transfer protocol
	OS				
	Windows Server 2003	2003	2003		Transfer protocol
	Audit and Control				
	Tripwire for Windows 2003	4.6.0			Transfer protocol
	Database platform				
	Clarion	??			
	Anti virus				
	Symantec Antivirus	circa 10			Transfer protocol
	Rendering software				
	TextPad	4.7			Transfer protocol
	Adobe Acrobat Reader	7			Transfer protocol
	Irfan View	3.98			Transfer protocol
	Open Office	2			Transfer protocol
	Microsoft Office Suite	2003			Transfer protocol

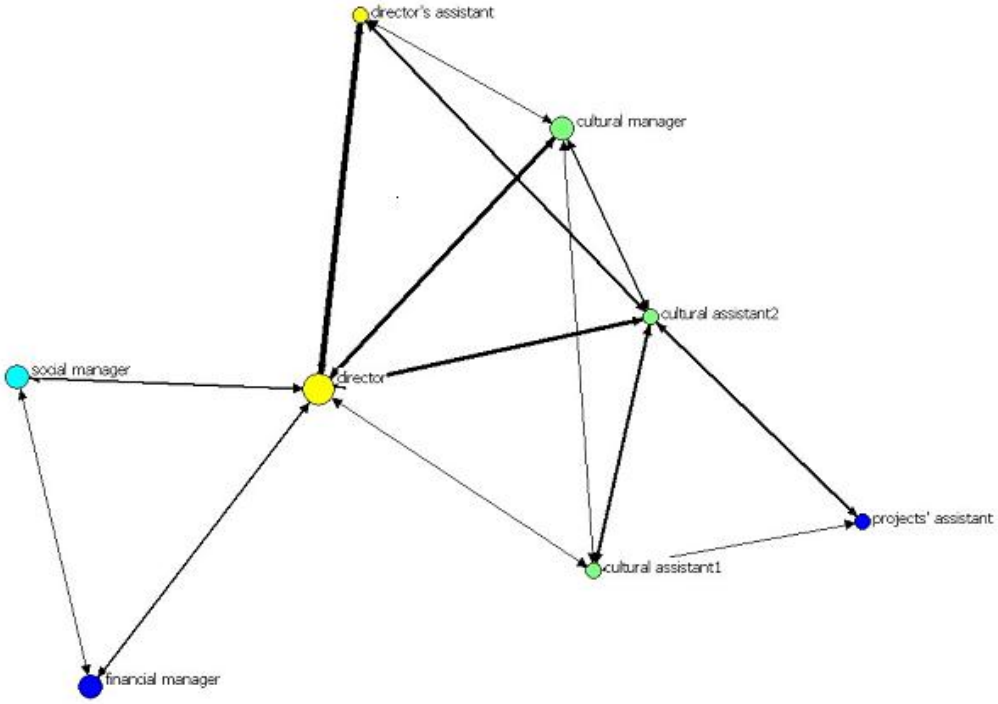
Appendix III: Network Diagrams

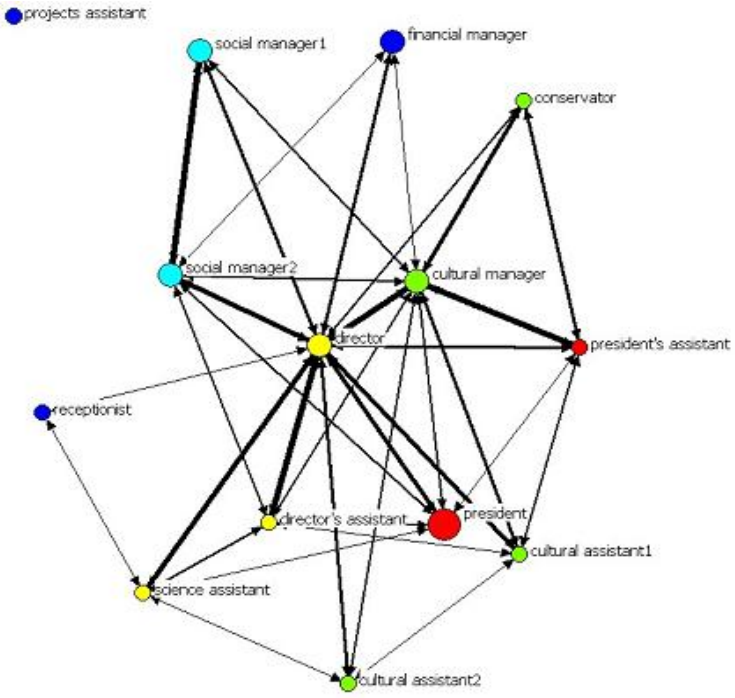
Timeline

1996
6 staff members
544 records

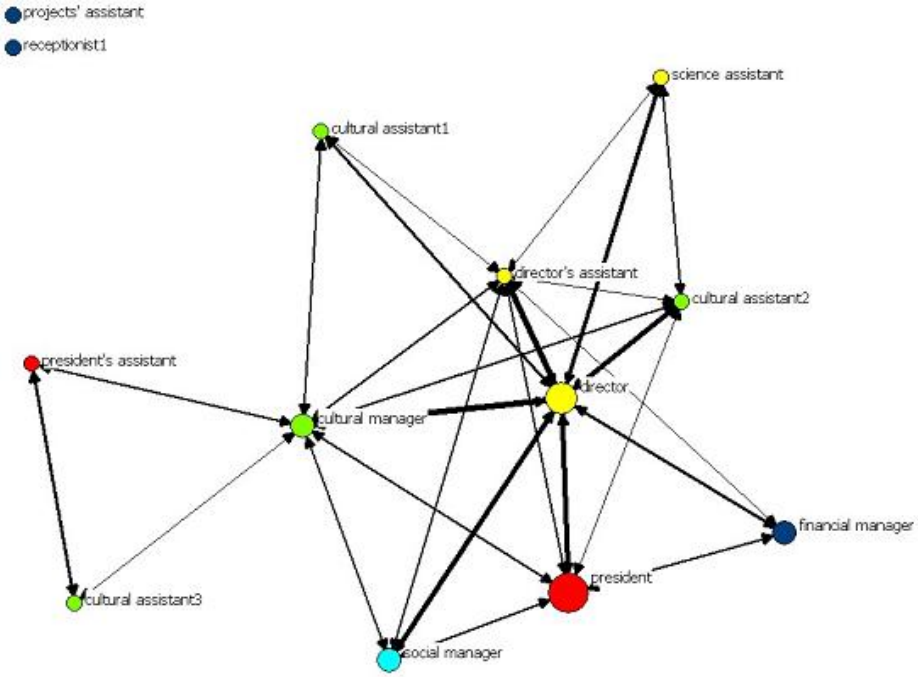


1997
8 staff members
712 records

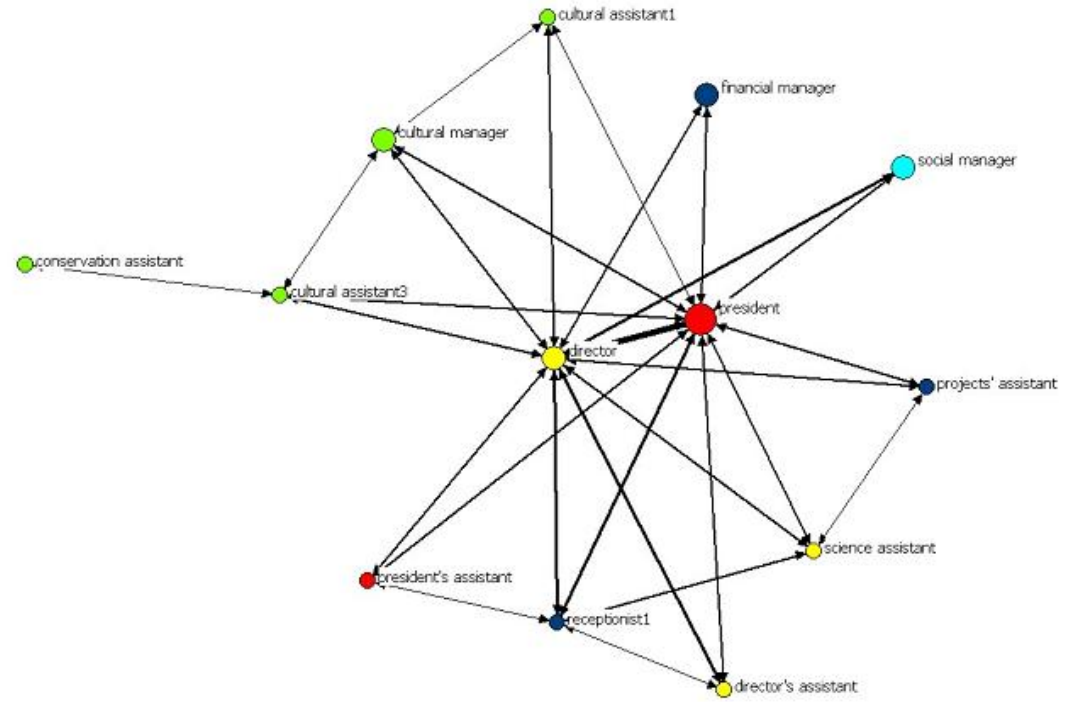




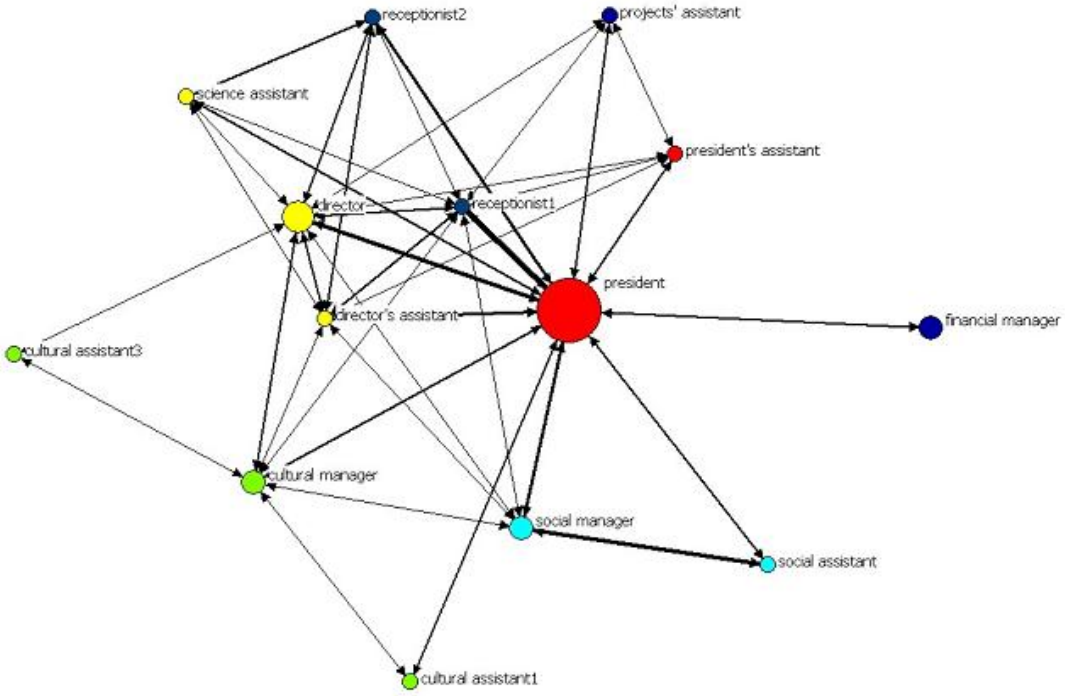
1998
14 staff members
1230 records



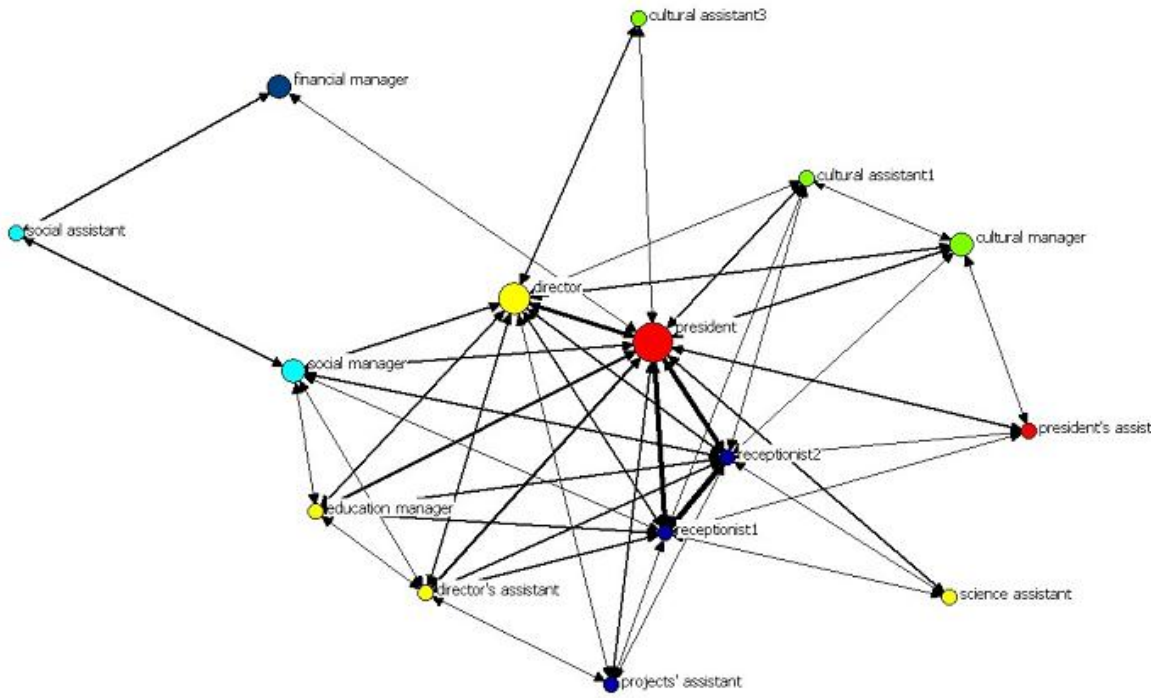
1999
14 staff members
1374 records



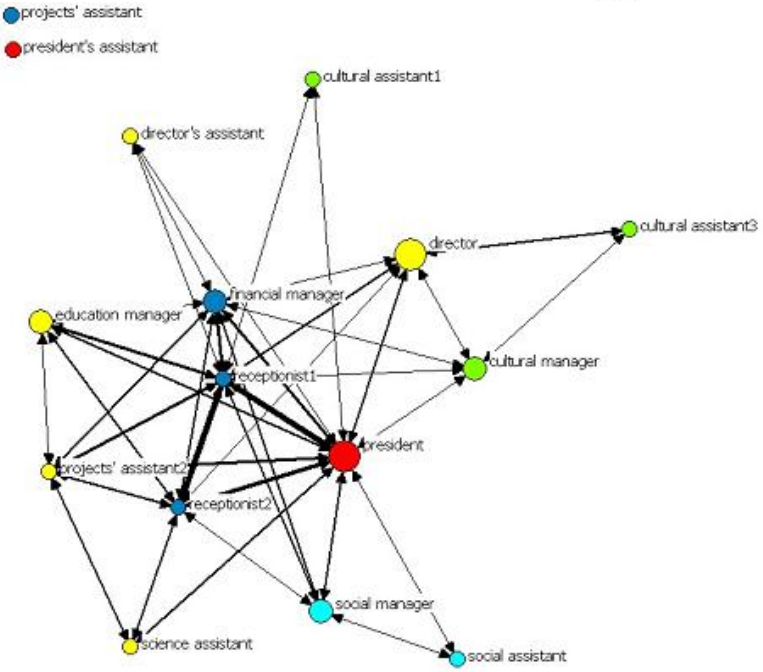
2000
13 staff members
1595 records



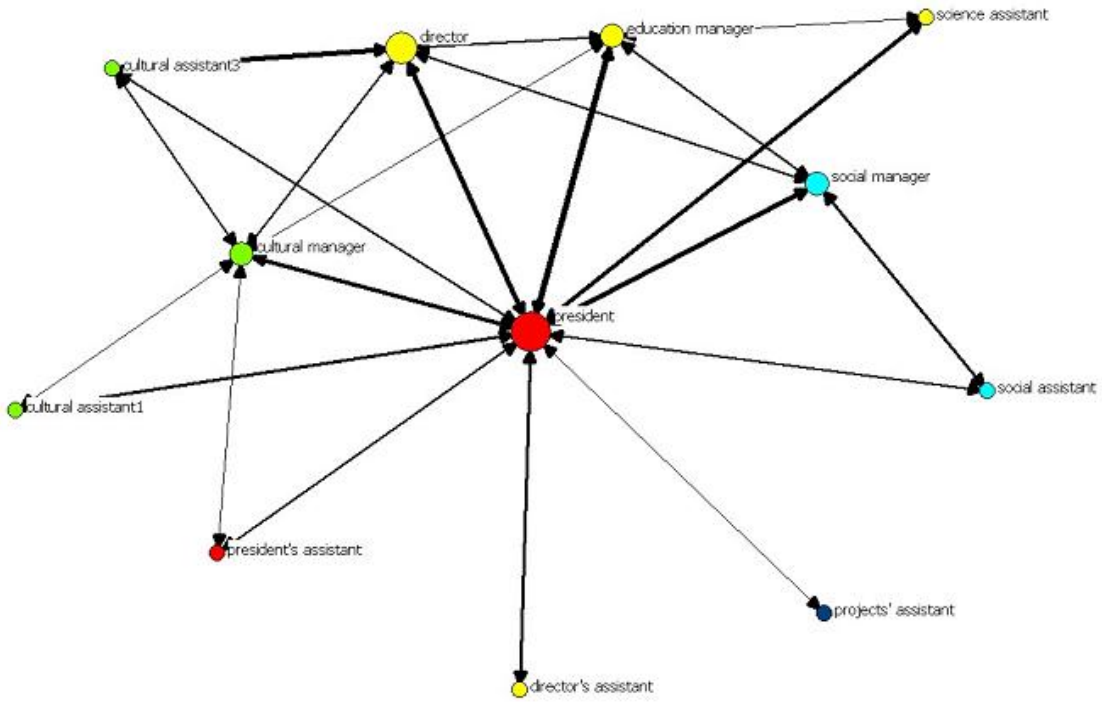
2001
14 staff members
2181 records



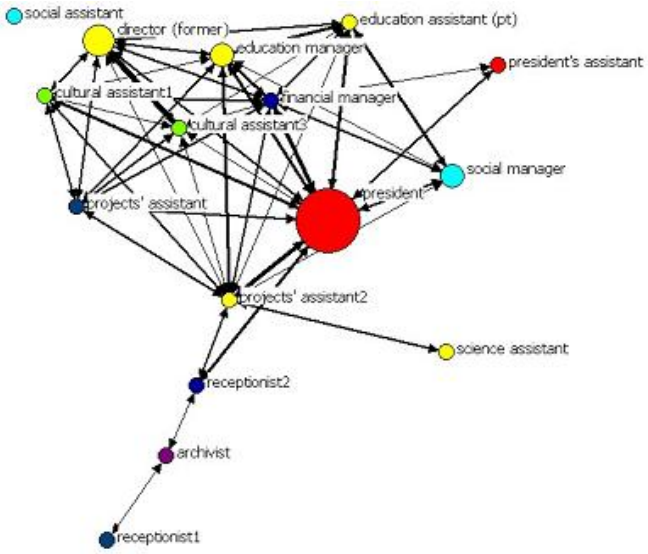
2002
15 staff members
2181 records



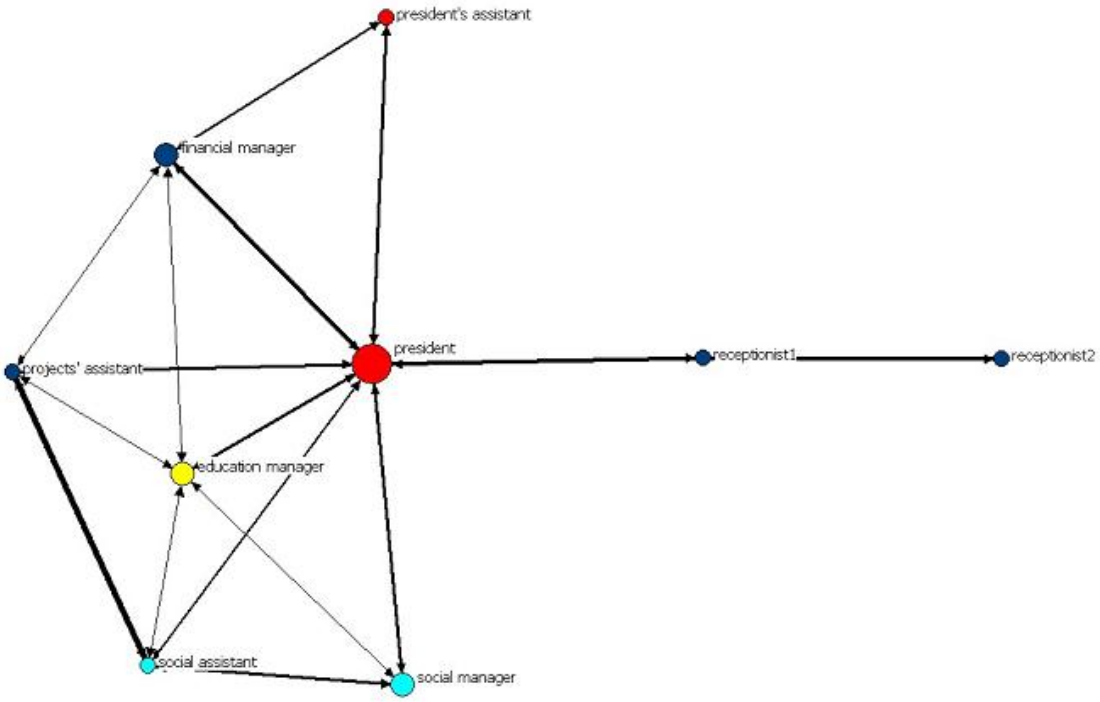
2003
16 staff members
3727 records



2003
Filtered view
11 staff members (4 filtered)
3617 records



2004
16 staff members
1667 records



2005
9
staff
members
707
records

Appendix IV: Transfer and Maintenance Protocol

The next is the translation from Spanish of the transference and maintenance protocol that I wrote in 2006 to instruct the IT consultants to perform the transfer of the contents of the networked server to the dark archive. The steps were accomplished in the next order:

- June-July of 2006: the server was purchased and prepared, file transfer and rendering were tested
- February 2007 - transfer
- July of 2007 – testing rendering of the databases, adjustments of the audit and control software

Transference of custody of the digital archive of Aleph Foundation

María Esteva
October 30, 2006

A preservation strategy for the digital archive

Aleph's digital archive will remain under a designated custodian during 10 years after the definite closure of the institution. The official closure date will be determined by the Inspectorate of Justice. The digital archive comprises all the digital contents currently located in the directory Dat on Rantor1 (G), (shared directory).The preservation strategy's goal is to maintain the archive "semi-functional" during the records retention period. This means that it will be possible to access all the electronic records (texts, images, Excel spreadsheets, etc.) and the financial and grant tracking databases, but it will not be possible to modify them. A preservation strategy comprised by three preservation measures was devised. The measures will be used in combination so that each one compensates the deficiencies of the other.[261]

- Copy of the bit sequence and migration of the operating system and creation /rendering software with the purpose of maintaining the records and databases functional during the next 10 years.
    - Copy or "backup" of the contents of Aleph's digital archive located in the current networked server to a - **"dark archive" -** new server hardware with Windows Server 2003 operating system
    - The dark archive will be loaded with current versions of the applications used to create and render the electronic records and databases.
    - The dark archive will be maintained turned-off under designated custody.
    - A computer screen and keyboard will be assigned to the server in order to perform maintenance operations and access the records and systems.
    - The dark archive's upkeep and the integrity and authenticity of its contents will be controlled once a year.
    - Annual verification of the need to migrate applications, equipment, and records.
- Technology preservation
    - Maintain the networked server in which the records are currently stored until it fails for mechanical reasons.
    - This server is defined from now on as **"static server."** [262]
    - The static server will be maintained tuned off under designated custody during the 10 years record retention period.
    - The server's functioning will be monitored once a year.
    - Upon the records retention period it can be used with the purpose of studying technology obsolescence in the future).
- Modified and multiple refreshing
    - Refreshing is, " the copying to another medium of a similar enough type that no change is made in the bit-pattern that is of concern to the application and operating system using the data."[263]
    - Copying should be done to a durable storage medium such as a DVD.
    - For security purposes the servers and the DVD copies of the archive should be kept in different sites.
    - Annual verification of the integrity of the back up support (currently DVD but the medium can change in the future) to determine migration.

Steps to follow:

The steps to archive the electronic records and maintain them accessible during the next 10 years are the next:

1. Preparation of the technical environment of the dark archive.
2. Preparation to copy the contents from the static server to the dark archive.
3. Transparent and audited transfer – copy – of the contents from the static server to the dark archive
4. Backup in DVD and annual monitoring of the backup media

5. Annual monitoring of the dark archive
6. Use of the records and systems in dark archive

Most of the steps will be completed by the IT consulting firm currently in charge of maintaining the systems at Fundación Aleph. Steps 1 and 2 have been partially completed already between them and me in the month of June. Transfer will be completed by the IT consultants. Evaluation that the transfer of the records and systems to the dark archive was done according to requirements will be done by me through the documentation generated and sent by the IT consultants. Once a year I will also supervise the step related to verifying the integrity of the back up media and determining the need to migrate software, hardware, or electronic records and databases by testing one of the backup copies of the contents of the dark archive.

1. **Preparation of the technical environment of the dark archive** (part of this work was accomplished during June and July of 2006)

- Updated versions of the software used until the foundation's closure to render the files as well as the source code and developers of the database systems (grant tracking and financial) will have dedicated spaces in the dark archive.
    - During July of 2006 the next software was installed: Microsoft Office 2003, Open Office, IrfanView, Adobe Acrobat, TextPad, Symantec AntiVirus, and the National Library of New Zealand Metadata Extraction tool. (See Metadata Timeline for versions of the software)
        - The IT consultant explained that only one version of Word can be installed in the server. Currently the version installed is Word 2003. I would like to explore the possibility of installing the versions of Word 95 and 97 currently installed in some of the office computers.
    - Eudora Pro should also be installed.
    - In the dark archive the rendering software should be located in a directory called `programasArchivo` separated from the directory that will contain the contents of the static server which is **EVERYTHING** that is currently located in the disk Dat on Rantor1 (G).)
- Inventory all the software applications in the dark archive and send me the list before the final transfer to determine if there is something else that needs to be installed.
- All the installation CDs of the programs should remain in a case and labeled and given to the designated custodian.
- The dark archive has been protected with an antivirus that I installed in June of 2006. It is important to update the version online before transferring the files.
- Install the software Tripwire for Windows
- Tripwire is an audit and control software. It can be configured to verify that files have not been modified and to control access to records and databases. Tripwire produces MDhashes of the indicated files and verifies changes against the initial

inventory. For this the settings that have to be implemented during installation are the next:
    - File adds delete modifications
    - Event tracking for objects
    - Last access time
    - Last write time
    - Create time
    - File type and size
    - Hash checking (This is the most important function because it will allow verifying if a change was performed to any file in the dark archive)
- To do the installation the IT consultant will have to learn how to use the program whose instructions are in English. I suggest Leonel for this task as well as to perform the annual verifications because he knows the foundation's systems and the configuration of the static server.
- Since Tripwire will be used during the next 10 years, Leonel has to write up the commands that he uses during the annual verification and leave the documentation with the designated custodian so that if he stops working with the IT consulting company, another person is able to perform the routine.

2. Preparation of the contents of the static server for transfer.

- All the passwords used to access the files and the database systems in the static server should be cleared before transfer. Especially those created to allow restricted access to certain directories and email back up boxes.
- All the databases should be locked so that no more data is entered.
- Perform an automatic inventory of the contents of the static server. This can be done through the command line and saved as a .txt file.
    - The inventory should contain the next elements:
        - File name
        - File path
        - Autor
        - Creation date, last modification date, last access date
        - File type
        - Size type
        - Ex. Retinaesp. 56KB Microsoft Word Document 6/26/2005 5:20pm 9/2/2005 6:09 PM.
- The file name should be `inventario-server-estatico-FA`
- The .txt file should be stored in a directory called `transferenciaArchivo` in the dark archive
- A copy of this file should be sent to me through email or in a CD to my postal address

3. Audited transfer

This step has two parts: a) transfer of the contents from the static server and b) transfer of limited access software (Programs of limited access are those used by certain staff members and that where not located in the shared directory).

a) Transfer of the contents of the static server

- The contents of the static server located on Dat on Rantor1 (G) have to be copied to a directory named `ArchivoAleph` in the dark archive.
- Copying should be done respecting the same virtual distribution in which the files are currently maintained, this is respecting the directories hierarchies.
- Transfer will be done with the program Total Commander. The transfer test performed indicates that the program is adequate for the task.
- Document the results of the transfer. When we tested the transfer we realized that Total Commander does not generate a report of the transfer. Results can be recorded by making a screen shot to document that there are no differences between the two directories (if there are differences the program will highlight those in colors).
- Generate an automatic inventory of the contents of the directory `ArchivoAleph` just like the one that was done in the static server.
- The file with the inventory should be called `inventario-server-archivo-FA`
- A copy of the inventories should be stored in the folder `transferenciaArchivo` en el servidor archivo (See fig.1).
- Use Tripwire to produce an MD5 *hash* of each one of the records in the directory `ArchivoAleph` in the dark archive
    - In the future, each time that a control is made on the server the hashes correspondent to each of the files should be compared to this first list.
- Verify the correct functioning of the grant tracking and the financial databases in the dark archive.
- Verify the correct functioning of all the rendering programs that have been installed in the dark archive such as: Word, Eudora, etc.
- Make shortcuts in the desktop to access the rendering software and the databases.
- Recap of the required documentation and procedure:
    - Narrative of the steps followed and their results
    - Transfer documentation (screen shots of the results)
    - Pre and post inventories
    - Results of the verification of the programs and databases
- Turn copies of these documents to the designated custodian on a CD
- Keep a copy of all the documentation in the directory `transferenciaArchivo` in the dark archive
- A copy of all the documentation should be sent to me by mail in a CD
    - With this documentation I will make a report confirming that the transfer was accurate.

b) Transfer of programs of limited access

- Software that was installed only in the computers of some staff members should be stored in the dark archive in a directory named `ProgramasAccesoLimitado.` These programs are:
    - TANGO, the human resources software located only in Lelia's work station
    - Lotus, in most secretaries and assistants stations, copy the version from the receptionist computer as it contains
- Make a list of the programs installed in this directory
- Create a shortcut from the desktop / interface to access them

4. Backup

- Make a minimum of three copies of the contents of the directory `ArchivoAleph` on DVD –R reading only.
    - This copies should be called `ALEPHArchivo` 1, 2, and 3 respectively
- Use a good quality DVD such as Fuji.
- The DVD should be adequately labeled so that their contents can be identified
    - Do not place any label nor write over the surface of the DVD.
    - Write the name of the archive and the date in the clear plastic circle of the center of the DVD with a permanent marker.
    - The DVD cases should be adequately labeled indicating the name, the recording date and the provenance (Fundación Aleph).
- A DVD should be kept by the designated custodian so they can be used for recovering purposes if that is needed.
- Another DVD should be sent to me. I will use this material to do the annual verifications and to determine whether the DVD has to be refreshed.
- The third DVD should be placed off-site in a secure place with controlled temperature and humidity such as the storage vaults offered by IRON MOUNTAIN to store electronic records.
- Other copies can be stored on a bank safe, or left under the custody of the institution's lawyers. The important thing is that they are distributed and to know who has them.

5. Annual and periodical controls

This step has the next parts: a) server's maintenance b) verification of the integrity of the records, c) documentation of every task, d) verification of need of software updates, migration of records, or backup refreshing.

a) *Maintenance of the dark archive*

- For security reasons the dark archive will not be connected to the Internet or Intranet

- The server will remain turned off. It will be turned on once a year for audit control purposes
- If the static server is kept, it should be stored in other location different from the dark archive.
- The password to access the dark archive will be known only to those authorized by Aleph to access the contents or to perform maintenance and verifications. The password should be kept in a secure place by the designated custodian along with the inventories, the instructions and the software installation CDs.
- A computer screen will be dedicated to the server in order to perform access and maintenance tasks.
    - Shortcuts to the rendering programs and to the grant tracking and financial systems should be created in the server's interface.
    - Create a shortcut to access the directory `ArchivoAleph`
- The routine check up of the server including the hardware and operating system, will be performed according to IT maintenance best practices.
    - The IT consultant will create a list of routines and procedures to be accomplished annually so that the employees do the same every year.

b) *Verification of the integrity of the records*

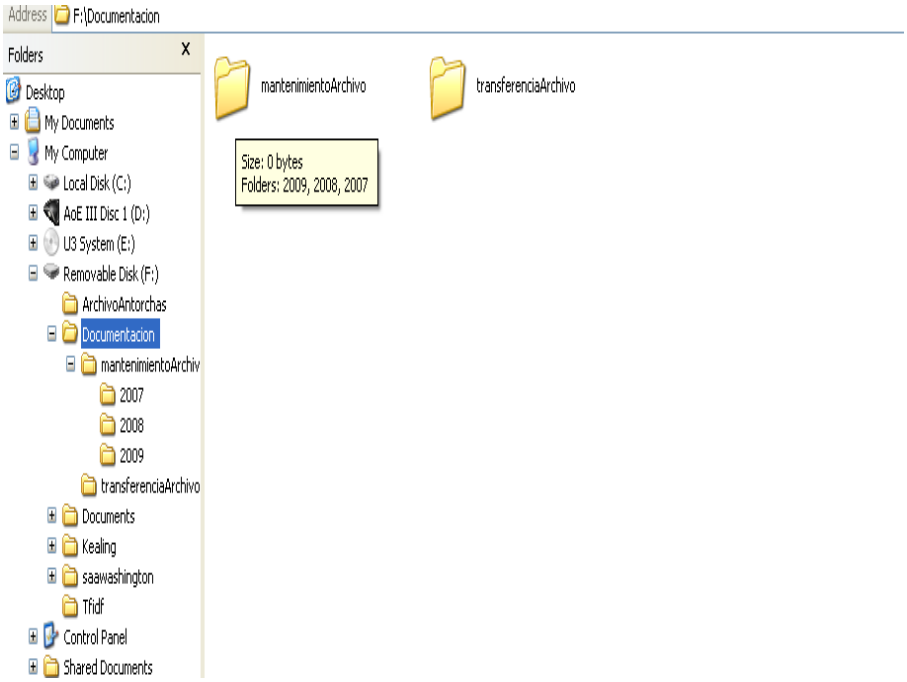
- Verification will be done once a year with Tripwire for Windows
    - The verification consists in generating a report after issuing a command to the Tripwire software to check the contents of the `ArchivoAleph` directory again the hashes saved during the first
    - The log/report will indicate if, when, who, somebody accessed and or changed the records.
    - Ideally the results of the verifications will be negative, indicating that no alteration was done.
    - Access to the records or databases will also be recorded
    - If a change in a record is verified, the record can be recovered from the backup DVD. Also, Tripwire can make automatic recoveries but since the server will be unplugged, automatic recoveries can't be programmed. When and if done, recoveries and the results should be documented.

c) Documentation

- The documentation generated after steps a) and b) will be stored in a directory called `mantenimientoArchivo` in the dark archive and within the directory labeled with the year correspondent to the date in which the maintenance is performed.
- The annual documentation consist on:
    - A report indicating that hardware and software maintenance was accomplishment and its results
    - The tripwire report on changes to records and access to the server

- This documentation will help identify patterns, incidents, and changes in records over time.
- These documents are what makes of the dark server an archive
- A copy of all the annual documentation should be sent to me for control.

Fig 1. Structure of the documentation directory in the dark archive



d) Verification of need of software updates, migration of records, and refreshing of back up media.

- Based on research about stability of media I will determine if the backup DVDs have to be refreshed and to which media.
- Using the back up DVD, I will verify records' rendering and functionality of the databases. Verification will be done by observing how a sample of records belonging to the different years represented in the archive render in newer versions of OS and software. In this way I will determine if there is need to update software or of migrating the actual records.
- If I detect the need to migrate records or software I will get in touch with Aleph's authorities and with the IT consulting firm to discuss the convenience/ feasibility of doing it.
- The results of these verifications will be recorded and sent as files to the designated custodian so that they are incorporated in the directory that contains all the documentation

6. Use of the dark archive

- Aleph authorities will determine who is authorized to access the archive and under what circumstances.
- Under all circumstances records and databases in the dark archive can only be read and not modified.
- Aleph should consider the future of the archive after the end of the records retention period of 10 years

Appendix V: File Rendering Testing Table

Record name	Last modified date	Format/version	Language	Word 2007/WindowsVista	Word 2003/Server 2003	Word 2003/Windows XP	Windows 98 Microsoft office 97	Notes
Date of test				Mar-08	Jul-06	Jul-06	Jul-06	
audit.doc	31/10/1991	Microsoft Word 5.0 for DOS	English	Legible, loss of certain code characters, bolded characters, en dashes and Spanish diacritics. Different glitches in header and footer. Squares and others not present	Legible, loss of: certain code characters, length distortion, bolded characters, en dashes, and Spanish diacritics	same as previous	same as previous	Opened with Windows encoding
marcos.doc	18/10/1992	could not be identified	Spanish	Legible with loss of Spanish diacritics but few differences in rendering for ex. ü shows up as a blank space in 2007 and as a square in 2003	Legible with loss of Spanish diacritics	same	same	Opened with Windows encoding
bgrdaprq.doc	8/30/1992	Microsoft Word 5.0 for DOS	English	Legible, character code loss in header and footer, loss of underlined, difference in glitches, squares are replaced by spaces in 2007 as well as other minor differences	Legible, character code loss in header and footer, loss in underlined titles	same	same	Opened with MS-DOS encoding

Record name	Last modified date	Format/version	Language	Word 2007/WindowsVista	Word 2003/Server 2003	Word 2003/Windows XP	Windows 98 Microsoft office 97	Notes
Date of test				Mar-08	Jul-06	Jul-06	Jul-06	
barracas.doc	7/22/1993	Microsoft Word 5.5 for DOS	Spanish	Legible, loss of certain code characters in header and footer and Spanish diacritics	Legible, loss of certain code characters in header and footer and Spanish diacritics	same	same	Opened with Windows encoding
conmus94.doc	8/20/1994	Microsoft Word 5.5 for DOS	English	Legible, loss of code characters in header and footer, minor glitches differ from 2003	Legible, loss of code characters in header and footer	same	same	Opened with MS-DOS encoding
museos95.doc	8/9/1995	Microsoft Word 5.5 for DOS	English	Legible, loss of code characters in header and footer, minor glitches differ from 2003	Legible, loss of code characters in header and footer	same	same	Opened with MS-DOS encoding
difvid.doc	26/6/1995	Microsoft Word 5.5 for DOS	Spanish & English	Legible, loss of code characters in header and footer and Spanish diacritics, minor glitches differ from 2003	Legible, loss of code characters in header and footer and Spanish diacritics	same	same	Opened with Windows encoding
freuden1.doc	8/10/1996	Microsoft Word 6.0/95	Spanish	Legible and formatted	Legible and formatted	same	same	opens normally
cmttbrsl.doc	1/8/1997	Microsoft Word 6.0/95	English	Legible and formatted	Legible and formatted	same	same	opens normally

Record name	Last modified date	Format/version	Language	Word 2007/WindowsVista	Word 2003/Server 2003	Word 2003/Windows XP	Windows 98 Microsoft office 97	Notes
Date of test				Mar-08	Jul-06	Jul-06	Jul-06	
frmnv985s.doc	22/10/1998	Microsoft Word 6.0/95	Spanish	Legible and formatted	Legible and formatted	same	same	opens normally
delarua.doc	26/10/1999	Microsoft Word 97 -2003	English	Legible and formatted	Legible and formatted	same	same	opens normally
mason02.doc	24/10/2000	Microsoft Word 97 -2003	English	Legible and formatted	Legible and formatted	same	same	opens normally
artargp.doc	18/10/2000	Microsoft Word 97 -2003	Spanish	Legible and formatted	Legible and formatted	same	same	opens normally
fcndzvr.doc	29/10/2001	Microsoft Word 97 -2003	Spanish	Legible and formatted	Legible and formatted	same	same	opens normally
reu2-09-19.doc	5/10/2002	Microsoft Word 97 -2003	Spanish	Legible and formatted	Legible and formatted	same	same	opens normally
eata01.doc	31/10/2003	Microsoft Word 97 -2003	Spanish	Legible and formatted	Legible and formatted	same	same	opens normally
schedu~1.doc	18/05/2004	Microsoft Word 97 -2003	Spanish & English	Legible and formatted	Legible and formatted	same	same	opens normally

REFERENCES

Part I

[1] Borges, J. L. 1980. “El Aleph.” In Prosa Completa, Vol. 2:112-125. Barcelona: Bruguera (original work published in 1949).

[2] The real name of the foundation is disguised for reasons of confidentiality.

[3] The archiving project actually concluded in April of 2008.

[4] The information contained in this section was obtained from annual reports issued by the organization and from interviews with the staff members. For reasons of confidentiality the references of the sources cannot be disclosed. This description is not intended as a historical summary about the foundation but as background information for the archival case study.

[5] This the foundation’s mission statement obtained from an annual report. It was included in the grant contracts and was publicly available through the institution’s website.

[6] While the operations concluded at the end of 2005, the foundation had to wait for the Inspectorate of Justice to officially determine the end of activities.

[7] Luis Priamo, email communication with the author, 12 March 2007.

[8] Taylor, G. 1996. *Cultural Selection*. New York: Basic Books.

[9] Cook, T. 2005. "Macro-appraisal in Theory and Practice: Origins, Characteristics, and Implementation in Canada, 1950-2000," *Archival Science*. 5: 101-61.

[10] In Argentina there is not yet a clear distinction between active files and archival files, both are called archives.

[11] Records management companies have been establishing in Argentina for less than ten years. First used by multinational companies, in the last three years they have begun to be used by smaller companies and by governmental organizations such as universities and tax agencies. As in the U.S, the business in Argentina improved following the September 11 attacks and the business scandals. See, Garcia Bartelt, M. "Los Escándalos Empresariales y el Ataque a las Torres Gemelas Impulsaron a Iron Mountain que Logró Crecer en Tiempos Violentos," *La Nación*, 2 March 2003; Kischner, N. "Cuidar los Documentos y el Negocio," *Revista Pyme Clarín*, October 1, 2005.

[12] Commercial records-management companies are not conceived as archives. Their profit stems from managing the records; this is from moving them, faxing them, shredding them when needed, etc. Most of the paper documents stored have shorter retention spans and are accessed by their creating organizations. This is why the bulk storage per se is not expensive, what is expensive is the actual management of the records. Instead, to store electronic media the vaults are considerably more expensive.

[13] These are rubricated ledger books into which the annual balances, the annual tax reports, and the board meeting minutes are copied on to and signed by legally authorized authorities. Except when they are sent to the copier, these books always have to remain in the institution and have to be presented during audits and inspections.

[14] From manuscript copying, to hard copies from originals, to copy-transfers, to digital impressions, the process of transferring information onto the legal ledger books has varied through the years. At the foundation, after the annual balances, the meeting minutes and the tax information was produced, the books and the information to be copied were taken to a company that did the transfer.

[15] The foundation had the practice of returning materials belonging to applicants who did not win awards.

[16] Rose, K. Assistant Director at the Rockefeller Archive Center, email communication with the author, 30 July 2004.

[17] Argentine Law 25.326, *Personal Data Protection Act*, 12 February 2003, [cited 21 July 200]. Available at: http://www.bcra.gov.ar/pdfs/marco/iHabeas%20Data.PDF

[18] *Code of Commerce of the Argentine Republic*, 1889, [cited 12 July of 2007]. Available at:http://www.pelaez.com.ar/pelaez/power/Otros/CodigodeComercio/CodigodeComercio.html

[19] The law does not specify the conditions or quality that those records should have in order to be acquired by the Argentine General Archive (AGN).

[20] The law 15.230 Archivo General de la Nación (AGN) was passed by Congress in 1961. It has an inconsistency in relation to the number of retention years for private corporate records stipulated by the Code of Commerce. The law quotes twenty years while the Code of Commerce states ten years after the company's cessation of activities is declared by the Inspectorate of Justice.

[21] Informal phone conversations that I conducted with personnel from the AGN in 2004 indicated that this is not a regular practice. If a private archive is acquired by the AGN it is because the archivists identify the fonds as valuable from a historical perspective or because creators/companies approach the archive to donate their records. In fact, I have participated in conservation projects involving state owned company archives or newspaper archives in which records of different types (photographic and architectural drawing materials) were discarded or rescued from the garbage bins by individuals because the AGN would not take them. I submitted the question of whether they receive requests from the Inspectorate of Justice to preserve certain archives to the AGN through email in April of 2008 and did not obtain a response.

[22] As of April of 2008, the foundation is waiting for the General Inspectorate of Justice to officially close its activities. This means that in the meantime it is obliged by law to present balances and maintain the legal and tax accounting books, even if they don't show any changes. It remains to be seen if the agency decides anything about the institution's archive.

[23] Control records refer to records that control the flux and path of the case files in a government record-keeping system, it can be paralleled to a registry. See "Tabla de Plazos Mínimos de Control de Documentos," Anexo Decreto 1571/81 *Ley del Archivo General de la Nación.* Ley 15930/61, [cited 30 July 2007]. Available from http://www.mejordemocracia.gov.ar/normativa/ARCHIVO%201/LeyNacionaldearchivos 15930.pdf

[24] The records management system in Argentine governmental institutions is not based on record groups or series. The record unit is called "expediente," a case file that carries all the records that pertain to each and every transaction, from the hiring of a staff member, a purchase, or the creation of a new governmental department. The tradition of the expediente goes back to 16th century Spanish recordkeeping practice.

[25] A shared directory is a virtual folder structure mapped on a networked file server or networked hard-drive which allows authorized members in a network to store, access, and share files.

[26] I worked for foundation Aleph from 1994 to 2000. During the first years I was under contract to conduct projects outside the organization and from 1997, when I became a full staff member, to the end of 1999 I worked in a conservation facility located in a different building and my work-records were in the facility's server. I opened a folder in the shared directory at Aleph in early 2000 when I moved to an office in the foundation's main building. Interestingly, when in 2004 the conservation facility was donated, the system's administrator transferred the contents of the networked server (located in that facility) to the networked server at Aleph. This confirms the inference that records were not tossed, but transferred from one server to the next.

[27] Emulation is a digital preservation strategy used to preserve the look and feel and the functionalities of a given program or file. The program or file to be emulated runs in a software or hardware that imitates the original platform. More information about the definition and characteristics of emulators can be found at: National Library of Australia, "Digital Preservation Strategies." *Preserving Access to Digital Information (PADI),* [cited 20 April 2008]. Available at http://www.nla.gov.au/padi/topics/19.html

[28] Clarion is a database development software. See SoftVelocity, *Clarion 7,* [cited 20 April 2008]. Available at:http://www.softvelocity.com/clarion/c6.htm

[29] This strategy is called "technology preservation" or "computer museum solution." See, Cornell University Library. 2003. "Digital Preservation Strategies," *Digital Preservation Management: Implementing Short Term Strategies for Long-Term Problems,* [cited 12 April 2008]. Available at: http://www.library.cornell.edu/iris/tutorial/dpm/terminology/strategies.html

[30] In September of 2004 I started my second year in the doctoral program.

[31] This assumption was based on previous work that I did during an Independent Study at the Austin History Center, processing the paper and electronic records of the Metropolitan Austin Independent Network (MAIN). In this case the same arrangement structure was applicable to both paper and electronic records, some of which were copies of others and some not. In May of 2004 I presented the paper "Processing MAIN Electronic Records" at the *Annual Meeting of the Society of Southwest Archivist*s in San Antonio Texas during the session "Changing Times: Changing Traditional Processing Practices."

[32] Pearce-Moses, R. and Davis, S eds. 2008. "Framework for Arrangement." *New Skills for a Digital Era,* [cited 10 April 2008], 6. Available from the Society of American Archivists (SAA) website at http://www.archivists.org/news/NewSkillsForADigitalEra.pdf

[33] DSpace is an open source institutional repository software developed by the Massachusetts Institute of Technology and Hewlett Packard. See DSpace, [cited 28 March 2008]. Available from http://www.dspace.org/

[34] The interviews covered the roles, tasks, relationships, and the technological aspects of the staff member's work. While I asked all the questions that I had planned, I also let them talk freely about what they remembered best or knew more about their work. As a result, I was able to use the answers for more than I had initially intended. After 2005 I traveled to Buenos Aires two more times and I could re-interview and corroborate information with some of the interviewees as needed later in the data analysis process. I also communicated through email with some of them. The interview protocol is included in Appendix I.

[35] The class Data Mining was taught in the fall of 2004 by Dr. Maytal Saar-Tsechansky from the Department of Information, Risk and Operations Management at the McCombs School of Business, University of Texas at Austin. Professor Saar-Tsechansky supervised my two semesters of independent studies in text mining.

[36] Paquet, L. 2000. "Appraisal, Acquisition, and Control of Personal Electronic Records: From Myth to Reality," *Archives and Manuscripts*, 71-91.

[37] An example are the instructions to conduct inventories of electronic records collections in unmanaged environments issued by the Public Records Office in the UK in which it is suggested that when the way in which records are organized is only understood by their author it might be appropriate not to include these records in the inventory. See Public Records Office. 2000. "Guidance for an Inventory of Electronic Records: a Toolkit," [cited 7 November 2007]. Available from the National Archives website at http://www.nationalarchives.gov.uk/documents/inventory_toolkit.pdf. In the chapter "Appraisal of Electronic Records" from the book *Thirty Ears of Electronic Records,* Linda Henry describes a similar reaction to emails that "typically include everything" concluding that they "can't appraise all the unorganized materials on two million personal computers because it has not been subjected to rigorous records management to both reduce the volume and arrange the substantive records into logical order." See, Henry, Linda J. 2003. "Appraisal of Electronic Records." In *Thirty Years of Electronic Records,* ed. B. I. Ambacher. Lanham, Maryland and Oxford: The Scarecrow Press, 38.

[38] The Association for Information and Image Management (AIIM) through their Enterprise Content Management Certificate Program offer courses to manage ROT. The statement included in this passage was sent through email on April 17 2008 as an advertisement of the course contents. For information about AIIM see, http://www.aiim.org/education/erm-content2.asp?id=30625

[39] In her presentation "Automating and Constructing Rules for Appraisal in the Digital Environment," Mariella Guercio described the contents of a shared directory with similar

characteristics to the one I describe in my study and mentioned that the way to appraise them is by analyzing records one by one. *Appraisal in the Digital World*, Accademia Nazionale Dei Lincei and DELOS NoE Conference, 15-16 November 2007, Rome, Italy.

[40] Eastwood, "Appraisal of Electronic Records," nd.

[41] The foundation had been a major supporter of the academic telecommunications agency.

[42] For a reference of archives as geological sediments see, Bautier, R. H. 1961. "Les Archives. L'histoire et ses methods." Paris: 1120. In L. Duranti .1994. "The Concept of Appraisal and Archival Theory." *American Archivist,* 335 57:328-344

[43] Hodder, I., ed. 2000. *Towards Reflexive Method in Archaeology: The Example at Catalhoyuk*. British Institute of Archaeology at Ankara Monograph No. 28. MacDonald Institute for Archaeological Research.

[44] When as a group we defined non-records, these were brochures, invitations to events, and miscellaneous materials that were not produced by the foundation nor related to its activities. For privacy issues we also decided to discard resumes of people soliciting jobs.

[45] Duranti, L. 1994. "The Concept of Appraisal and Archival Theory." *American Archivist* 34, 57: 328-344. In this article Duranti states that records creators should not be made aware of the power of the documents that they create because then they will try to alter them.

[46] National Archives, UK. *Digital Record Object Identification. (DROID)*, [cited 12 January 2008]. Available at: http://droid.sourceforge.net/wiki/index.php/Introduction

[47] FileMerlin is a commercial file format conversion tool that in the process of converting the files identifies them. See, Advanced Computing Innovations, Inc. *FileMerlin, Advanced File Conversion Software*, [cited 8 February 2008]. Available at http://www.acii.com/fmn.htm

Part II

[48] InterPARES 2. 2002. *Case Studies*, [cited 5 March 2008]. Available from the InterPARES website at http://www.interpares.org/ip2/ip2_case_studies.cfm

[49] M. B. Schiffer. 1996. "Pathways to the Present: In Search of Shirt-Pocket Radios with Subminiature Tubes." In *Learning from Things: Method and Theory of Material Culture Studies*, ed. D. Kingery. Washington and London: Smithsonian Institution Press, 81-88.

[50] M. B. Schiffer. 1996. “Formation Processes of the Historical and Archaeological Records.” In *Learning from Things: Method and Theory of Material Culture Studies*, ed. D. Kingery. Washington and London: Smithsonian Institution Press, 74.

[51] Lubar, S. 1996. “Learning from Technological Things” In *Learning from Things: Method and Theory of Material Culture Studies*, ed D. Kingery. Washington and London: Smithsonian Institution Press: 31-35; Corn , J. 1996. “Object Lessons/Object Myths? What Historians of Technology Learn from Things.” In *Learning from Things: Method and Theory of Material Culture Studies*, ed D. Kingery. Washington and London: Smithsonian Institution Press, 35-54.

[52] See definition of authenticity in Pearce-Moses, R. “A Glossary of Archival Records Terminology.” *Society of American Archivists* [cited 6 March 2008]. Available from the Society of American Archivists website at: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=9 .

[53] Kirschenbaum, M. 2001. “Materiality and Matter and Stuff: What Electronic Texts Are Made Of.” In *Electronic Book Review* [electronic bulletin board], [cited 23 January 2008]. Available at http://www.electronicbookreview.com/thread/electropoetics/ .

[54] Rekrut, A. 2005. “Material Literacy: Reading Records as Material Culture.” *Archivaria*, 60: 11-35.

[55] SoftExperience, *Metadata Miner*, [cited 23 January 2008].Available from: http://peccatte.karefil.com/software/Catalogue/MetadataMiner.htm .

[56] .prg and .dbs are dBase file types.

[57] The seamless transfer of data was confirmed by two different IT consultants who designed the second and third systems for the foundation. dBase data is easy to convert because it contains ASCII coded data and the record format indicates where each variable in the database starts and ends. Also, the entire database history for Aleph was under the relational database model which allowed the migration to occur without major problems.

[58] Carlos, interview, 2005. Carlos was the systems’ administrator at the foundation from 1998 to 2005.

[59] Information about computing equipment purchases was obtained from the foundation’s accounting books.

[60] These reasons where extracted from the interviews to different staff members.

[61] This was expressed during my interview with one of Renee's first assistants Lara. Here, the use of the term archive corresponds to the common use of the term in Argentina at the time, both in reference to records that are filed in the creating organization as well as those stored in archival institutions outside of the creating organization. The term records management system has only recently been introduced, and it is not a familiar concept to many people. Since 2000 and after a curriculum reform, students in the School of Archival Studies at the Universidad Nacional de Córdoba can take a course in records management.

[62] Elena, interview, 2005.

[63] Pedro, interview, 2005.

[64] This is clear from the paper records' inventories finished in 2006. The bulk of records found in most of the offices date from the time in which the last new staff member in the area was hired and reflect the record-keeping practice that he or she implemented. Records prior to 1993 are filed in the centralized file created by Renee and stored in the basement.

[65] Natalia, interview, 2005.

[66] Eliana had also been an executive secretary before she arrived at Aleph and knew the filing standards.

[67] Administrative and personnel records were subjected to audits and to regulations issued by the General Inspectorate of Justice, the tax agency ANSES, and the Ministry of Labor. These records had from the beginning of the institution their own set of filing practices, although those were somewhat modified with changes in staff members. Of all the documentation generated in the area, only the receipts were discarded after ten years. These were filed in chronological order so it was easy to dispose of them when their retention date expired.

[68] During the process of inventorying the foundation's paper records, accomplished from 2004 to 2006, I found that the institution is thoroughly documented. For example, we found a file with the lists of attendees and the menus served in the exclusive luncheons offered at the foundation from 1985 to 1998. This file was maintained by the cook who worked during that period. We were happy that nobody thought about throwing it away.

[69] According to accounting records, the institution bought five electric typewriters between 1985 and 1986.

[70] Victor, interview, 2005.

[71] From the interviews it was not possible to determine the precise dates in which the transition took place. Inferences had been made from purchasing data from the accounting books and from the presence of early files in the networked server.

[72] These were AT and 386 PC clones for the period 1991 to 1993, and 486 PC clones through 1995. The first upgrade to Pentium was purchased in 1996.

[73] This information was obtained from the foundation's Annual Report 1994-1995.

[74] Pedro, interview, 2005.

[75] See Appendix II Metadata Timeline.

[76] Transcribed verbatim of the interview with Pedro, 2005.

[77] Victor, interview, 2005.

[78] This comment was ascertained by all the interviewees that worked at the foundation at the time.

[79] When interviewed, the IT consultant who created the first system mentioned that he started working on it in 1986. The modification dates in program files corresponding to the system and found on the networked server show that in 1987 the system was in the development process.

[80] Pedro, interview, 2006.

[81] Larsen, M. D. 1985. "The Emergence of Microcomputers in Latin America." *Hispania*: 873-876. Available through JSTOR.

[82] The first time I used a PC was in 1989 as an intern at the Library of Congress in the US. At the National Library were I had been an employee since 1987 and until 1989 there were no computers. When I came back to the National Library in 1991 there were no computers either.

[83] Victor, interview, 2005.

[84] Larsen, "Emergence of Microcomputers," 1985.

[85] Pedro, interview, 2005.

[86] First IT consultant that worked for Aleph, interview, 2004. Novell NetWare 286 was released in 1985. The product, a file sharing system that could be used in a network of PC's, emerged in 1983.

[87] Written for dBase III, the routines found in the server are human readable. They reveal that projects could be discriminated by area (Culture, Social, Sciences and Education) and by various periods of time (year, month) as well as by whether they were approved or not. The database had searching and editing capabilities. The financial system controlled payments, produced statistics including amounts stipulated for each area, and listed date and type of accounting routines.

[88] Causey, J. "Netware." In *High Performance Network Unleashed,* [cited 2 April 2008]. MacMillan Computer Publishing. Available at: http://docs.rinet.ru/NeHi/ch22/ch22.htm .

[90] Within the directory that contains the system's files, one consists of all the Word documents with letters of acceptance.

[91] The first website of the foundation harvested by the Internet Archive dates from December of 1996, the same year in which the Internet Archive started. It is likely that this was the year in which the foundation launched its web presence. Files with content to post on the web found in the shared directory date from 1996.

[92] Due to non-disclosure issues, I cannot reveal the websites URL.

[93] This was probably since 1997 when the foundation purchased a Sun Microsystems Internet server. The server was located in the telecommunications agency site.

[94] After 2002 computer equipment updates were rare. One station was upgraded to Pentium IV and only a handful of Windows 2000 operating systems licenses were purchased.

[95] Pedro, interview, 2005.

[96] Ibid.

[97] The next quotes belong to interviews with Elena, Natalia, Cynthia, and Eliana, 2005.

[98] Pedro, interview, 2005.

[99] Cynthia, interview, 2005.

[100] Diana, interview, 2005.

[101] Pedro, interview, 2005.

[102] The source for this section was the interview with Pedro, 2005.

[103] The career of scientific computer specialist (computadora científica) was the first formal university level program in computer science. It was established in 1963 and had shorter duration than other degrees as it was intended as an aid to scientists using computers and to companies that were starting to use computers as well. See Jacovkis, P. 2004. "Breve Resumen de la Historia de la Computación en Argentina," [cited 23 March 2008].Sociedad Argentina de Informática. Available from SADIA website at: http://www.sadio.org.ar/modules.php?op=modload&name=News&file=article&sid=50 .

[104] Elena, interview, 2005.

[105] See CERP, Collaborative Electronic Records Project, Smithsonian Institution and The Rockefeller Archive Center. 2007. "Retention and the Digital Archive." In *Email Guidance*, [cited 2 February 2008], 4-6. Available at: http://siarchives.si.edu/cerp/progress.htm#guidance. The section briefly describes attitudes towards email retention that match the practices followed by the staff members at Aleph regarding their electronic text records.

[106] Pedro, interview, 2005.

[107] According to Luciana Duranti, InterPARES project director, data sets are recently considered as records by archivists as a consequence of studies of electronic record-keeping systems and how scientists consider their data as records. See, Duranti, L. 2005. "The Long-Term Preservation of Accurate and Authentic Digital Data: the InterPARES Project." *Data Science Journal*, 4:106-118.

[108] Jacovkis, *Breve resumen*, 2004.

[109] Sometimes corrections were done on the paper version and amended in the electronic one, or final documents were compiled on paper and not electronically.

[110] In the Culture area there was one yearly call for fellowships and grants, and the area of Science and Education called for research grant proposals in June and for fellowships in November of each calendar year.

[111] Within the grant tracking systems folders in the networked server there are thousands of .doc files corresponding to the letters of offer issued by the system and edited by the project coordinators and assistants.

[112] Eliana worked from 1995 until 2006 always as assistant to the president.

[113] The same sudden awareness happened with another interviewee who also went to her computer to remember if she followed some kind of pattern in her electronic record-keeping system.

[114] Jose, interview, 2005.

[115] Natalia, interview, 2005.

[116] Cynthia, interview, 2005.

[117] Ibid.

[118] The comparison between the paper and the electronic records organizational structure is in this dissertation is succinct and based on the inventories and description of the series and sub-series done to transfer the paper archive to the records management company. In the future, and given access to the paper files, a comparative study can explore the way in which a hybrid archive (paper and electronic records) relates, complements, and overlaps.

[119] It is not possible to include an appendix with the series and sub-series structure because the titles will disclose confidential information about the organization.

[120] Natalia, interview, 2005.

[121] It has to be noted that reports of external evaluations made by domain experts to the science and education area programs are present in different versions as well as in English and Spanish as electronic records in the shared directory.

[122] I found a definition of hybrid records systems in the site of the American Health Information Management Association. The definition is, "A hybrid health record is a system with functional components that include both paper and electronic documents and use both manual and electronic processes." See, AHIMA. 2003. "The complete medical record in a hybrid EHR environment Part III: Authorship of and Printing the Health Record, (AHIMA Practice Brief), [cited 12 April 2008]. Available from the AHIMA website at: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_021583.hcsp?dDocName=bok1_021583

[123] For an example on regulated use of email in the work-place see, Government of Australia. 2000. "Guidelines on Workplace Email, Web browsing and Privacy," [cited 10 April 2008]. Office of the Privacy Commissioner. Available at: http://www.privacy.gov.au/internet/email/ More regulations appeared as a consequence of the ENRON and Martha Stewart business scandals.

[124] Baron, J. 2007. *The Lawyer's Nose Under The Archivist's Tent: Appraisal Under Attack in the Age of Lawsuits, Search Engines, and The Quest for Total E-Record Archiving.* Paper presented at the Accademia Nazionale Dei Lincei and DELOS NoE Conference. Appraisal in the Digital World Accademia Nazionale Dei Lincei. Rome, Italy, November 15–16.

[125] InterPares.2002. Authenticity Task Force, *Requirements for Assessing and Maintaining the Authenticity of Electronic Records*, [cited 14 February 2008]. Available from: http://www.InterPARES.org/_file.cfm?doc=ip1_authenticity_requirements.pdf ; Bearman, D and Trant, J. 1997. "Electronic Records Research Working Meeting May 28-30, 1997: A Report from the Archives Community." *D-Lib*, [cited 14 February 2008]. Available from http://www.dlib.org/dlib/july97/07bearman.html; Cox, R. J. 1997. "Electronic Systems and Records Management in the Information Age: An Introduction" *ASIS* : 23(5), [cited 14 Feburary 2008]. Available from http://www.asis.org/Bulletin/Jun-97/cox.html ; Duranti, L et al, .2002. *Preservation of the Integrity of Electronic Records.* Boston: Kluwer Academic; R. Jones R. et al. 2006. *The Institutional Repository*. Oxford, UK: Chandos Publishing (Oxford) Limited.

[126] Cronenwett, P. L. 1984. "Appraisal of Literary Manuscripts." In *Archival Choices: Managing the Historical Record in an Age of Abundance*, ed. N. E. Peace. Lexington, MA: Lexington Books, 105-106.

[127] Jenkinson. *A Manual of Archive Administration*, 1937.

Part III

[128] Eastwood, T. 1992. "Towards a Social Theory of Appraisal," In *The Archival Imagination: Essays in Honor of Hugh A. Taylor*, ed. B. L. Craig. Ottawa: Association of Canadian Archivists, 71-89.

[129] Cook, T. 1992. "Mind Over Matter: Towards a New Theory of Archival Appraisal." In *The Archival Imagination: Essays in Honor of Hugh A.Taylor*, ed. B.L.Craig. Ottawa: Association of Canadian Archivists, 38-70.

[130] Ham G. F. 1993. "Appraisal Theory and Selection Goals." Chap 2 in *Selecting and Appraising Archives and Manuscripts* Chicago: The Society of American Archivists, 7-8 ; Schellenberg, T.R. 1956. *The Appraisal of Modern Public Records, Bulletin of the National Archives No.8*. Washington D.C: National Archives, 1-46.

[131] Henry, "Appraisal of Electronic Records," 2003.

[132] Nesmith, T. 2002. "Postmodern Archives: The Changing Intellectual Place of Archives." Lecture presented at the meeting of the Society of American Archivists.

[133] Taylor, H. 1987-1988. 2003 “Transformation in the Archives: Technological Adjustment or Paradigm Shift?” In *Imagining Archives: Essays and reflections by Hugh A. Taylor*, eds. T. Cook and G. Dods. The Society of American Archivists and The Association of Canadian Archivists.

[134] Samuels, H. 1986. “Who Controls the Past?” *American Archivist*, 49:109-124; Samuels, H.1998. *Varsity Letters: Documenting Modern Colleges and Universities*. Chicago: The Society of American Archivists.

[135] Cook, T. 2001. *Appraisal Methodology: Macro-Appraisal and Functional Analysis, Part B: Guidelines for Performing Archival Appraisal on Government Records,* [cited 4 November 2007]. Available from the National Archives of Canada website:http://www.archives.ca/06/061102_e.html

[136] Boticelli, P. 2000. “Records Appraisal in Network Organizations.” *Archivaria,* 49:161-191.

[137] Duranti, L. 2001.“The Impact of Digital Technologies on Archival Science.” *Archival Science*, 1: 39-55.

[138] Cox, R. J. 1997. “Electronic Systems and Records Management In the Information Age: An Introduction.”.*ASIS: 23*(5), [cited 3 December 2007]. Available at: http://www.asis.org/Bulletin/Jun-97/cox.html

[139] Information about this project can be located from the recovered site at: http://www.archimuse.com/papers/nhprc/

[140] Duranti, L. et al. 2002. *Preservation of the integrity of electronic records*. Boston: Kluwer Academic Publishers.

[141] Department of Defense, DoD 5015.2 STD. 2002. *Design Criteria Standard for Electronic Records Management Software Applications*, [cited 3 December 2007]. Available at: http://www.dtic.mil/whs/directives/corres/pdf/501502std.pdf

[142] Eppard, P. and Gilliland-Swetland, A. 2000. “Preserving the Authenticity of Contingent Digital Objects: The InterPARES Project.” *D-Lib* 6:7/8, [cited 29 April 2008]. Available from: http://www.dlib.org/dlib/july00/eppard/07eppard.html

[143] InterPARES Authenticity Task Force .2002. *Requirements for Assessing and Maintaining the Authenticity of Electronic Records*, [cited 4 January 2008]. Available from InterPARES website at: http://www.InterPARES.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf

[144] Brothman, B. 2002. “Afterglow: Conceptions of Records and Evidence in Archival Discourse.” *Archival Science*, 2:311-342.

[145] InterPARES Authenticity Task Force. 2005. *Authenticity Task Force Report*, [cited 6 October 2007]. Available from InterPARES website at: http://www.InterPARES.org/display_file.cfm?doc=ip1_atf_report.pdf

[146] Jenkinson, H. 1937. *A Manual of Archive Administration*. London: Percy Lund, Humphries & Co.

[147] Duranti, “The Concept of Appraisal,” 1994. In this article in note 20, page 334. Duranti states, “Impartiality is a characteristic of archival documents, not of their creators, who are naturally partial to their own interests.”

[148] This article emphasizes that records can be purposely biased. See, O’ Toole, J. 1999. “Cortes’s Notary: The Cultural Meanings of Record Making.” *Research Libraries Group. Annual Membership Meeting*, [cited 14 April 2008]. Available at:http://www.rlg.org/annmtg/otoole99.html

[150] Perer A. et al. 2006. “Using Rhythms of Relationships to Understand Email Archives.” *JASIST*, 57, 14: 1936-1948.

[151] Leuski et al. 2003. “eArchivarius: Accessing Collections of Electronic Mail.” *Proceedings of the 26th Annual International ACM SIGIR*. Toronto, Canada.

[152] Heer, J. 2004. *Exploring Enron: Visualizing ANLP Results*, [cited 20 November 2007]. Available at: http://jheer.org/enron/v1/

[153] Sinclair, S. and Ruecker S. 2006. *Mandala Rich Prospect Browser*, [cited 7 June 2007]. Available at: http://mandala.humviz.org/

[154] Boticelli, “Records Appraisal in Network Organizations,” 2000.

[155] Duranti, L. and Guercio, M. 1997. “Research Issues in Archival Bond. Electronic Records Meeting, Session I.” *Archives and Museum Informatics*, [cited 29 March 2006]. Available from Archives and Museum Informatics website at: http://www.archimuse.com/erecs97/s1-ld-mg.HTM

156 Jenkinson, *A Manual of Archives Administration*, 1937.

[157] The combination of tools and techniques involved in this method was suggested by Dr. Maytal Saar-Tsechansky, currently an Assistant Professor in the Department of Information and Risk Management in the McCombs School of Business at the University of Texas at Austin. To my best knowledge, this combination had never been implemented nor tested before I developed it and put it in practice.

[158] Hearst, M. 2003. *What is Text Mining?*, [cited 4 January 2008]. Available at: http://people.ischool.berkeley.edu/~hearst/text-mining.html. In this explanatory essay, Hearst specifically points to the difference between text mining and information retrieval. He also makes a distinction between text mining and information extraction which is closer to the type of work currently done with email corpora.

[159] While the shared drive contains records dating from 1991, the small amount of staff members and texts present from 1991 to 1996 made them irrelevant for the purposes of this study.

[160] FileBoss is a file management software. It took research and testing to find the right program to perform the many tasks needed to pre-process these texts and specifically to build the sets. FileBoss V2, [cited 23 February 2008]. Available at: http://www.theutilityfactory.com/

[161] Advanced Computing Innovations, Inc. *FileMerlin, Advanced File Conversion Software,* [cited 23 February 2008]. Available at:http://www.acii.com/fmn.htm

[162] Corpora List is an international list for consultation about text corpora. Available at: http://gandalf.aksis.uib.no/corpora/

[163] Slaton G et al. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM*, 18, 11: 613–620. This was the first article published about the vector space model.

[164] Baez-Yates, R. and Ribeiro-Neto, B. 1999. "Vector Model." In *Modern Information Retrieval*. New York: ACM Press, 27.

[165] Belew, R. K. 2000. "3.4 Vector Space." In *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge University Press

[166] McCallum, A. K. 1996. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*, [cited 5 February 2008]. Available at: http://www.cs.cmu.edu/~mccallum/bow

[167] Refers to the ISO/IEC 8859-1, a standard character encoding of the Latin alphabet.

[168] A Spanish stemmer can be downloaded from: Oleander Stemming Library, [cited 5 February 2008]. Available at: http://www.oleandersolutions.com/stemming/stemming.html

[169] I found this phrase in the fortune cookie during the time in which I was working on this piece of the research.

[170] Garcia, E. 2005. "Information Retrieval Tutorial." *Mi Islita,* [cited 4 February 2008]. Available from http://www.miislita.com/information-retrieval-tutorial/indexing.html

[171] Salton, G. and Buckley, C. 1988 "Term-weighting Approaches in Automatic Text Retrieval." *Information Processing & Management,* 24(5): 513–523.

[172] Sullivan, D. 2001."What is Text Mining? In *Document Warehousing and Text Mining*. New York: Wiley, 323-368.

[173] All the formulas explained and shown in this section were developed and adjusted to the characteristics of the data by Dr. Hai Bi.

[174] This approach was suggested by Dr. Maytal Saar-Tsechansky after my concern that averages, depending on the strength of similarity of the records involved in the calculation, could diffuse the relationship between two people in which one has many more records than the other. I found that the sum formula can provoke (again depending on the strength of similarity) the opposite effect and highlight strength based on quantity more than on similarity.

[175] Both the stemmer function and the filtering function were programmed as optional. The first one in Rainbow and the second one in the program developed for the dissertation.

[176] For the purpose of this dissertation the initials of the staff members are disguised.

[177] This phrase was said during a presentation of Johanna Drucker's work at the Roundtable Panel: "Modeling and Visualizing Historical Narrative." *Digital Humanities 2007,* June 2007, University of Illinois, Urbana-Champaign.

[178] Tichy, N. M. et al. 1979. "Social Network Analysis for Organizations." *Academy of Management Review*, 4(4):507–519.

[179] Hanneman R.A and Riddle M.2005. "Social Network Data." Chapter 1 Introduction *to Social Network Methods*: Riverside California: University of California, Riverside [cited 5 March 2008]. Available at: http://www.faculty.ucr.edu/~hanneman/nettext/

[180] Cross, T et al. 2002. "Making Invisible Work Visible: Using Social Network Analysis to Support Human Networks." *California Management Review,* 44(2):25–46.

[181] Analytic Technologies. UCINET 6: Social Network Analysis Software, [cited 3 March 2008]. Available at: http://www.analytictech.com/downloaduc6.htm

[183] ParaView, Parallel Visualization Application, [cited 21 February 2008]. Available at: http://www.paraview.org/New/index.html

[184] Horwitz, V. Email communication with the author, November 2007.

[185] Visualization results were also validated through qualitative interviews in the project: Perer. A and Smith M. 2006. "Contrasting Portraits to Email Practices: Visual Approaches to Reflections and Analysis." *Proceedings of the Working Conference on Advanced Visual Interfaces,* Venezia, Italy, 389-395.

[186] The use of distribution curves to validate the average results was suggested to me by Dr. David Dubin, Research Associate Professor at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign.

[187] Eliana did not remember why her records appear in the shared drive only from 1998 nor if she started using computers for the first time around that time, although she suspected it was at an earlier date.

[188] In Appendix III it can be observed that the closeness of the receptionist to the center of the network begins in 2001, when the foundation stopped sending the calls for grants through regular mail. Before that, the receptionists are always in the outskirts of the network and mostly in relationship with the science assistant.

[189] From the narrative of the archives formation process.

[190] See Appendix III containing the network diagrams for the ten years. Through the years, the cultural assistant has few above average relationships but consistently with the cultural manager.

[191] See Appendix III with the network diagrams for the ten yearly sets. Throughout her tenure, the position of the cultural assistant 1 in the outskirts of the network is a constant. This coincides with her habit of keeping many records including those sent by beneficiaries which interpreted from a social network analysis perspective implies her connection with the outside.

[192] First I tested removing only the receptionists' records, but since the records of the financial manager and the projects' assistant 2 were creating the same kind of tension I decided to remove them as well.

[193] This is well reflected in the animated visualization that presents the director's relationships with the rest of the staff members over a period of nine years.

[194] Gini, C. 1921. "Measurement of Inequality and Incomes." *The Economic Journal,* 31: 124-126.

[195] Patricia Galloway, during a class "Lifecycle of Metadata Objects" that I took in 2003 discussed how electronic archives could be presented in different arrangements.

[196] Leuski, *eArchivarious*, 2003.

[197] Dr. Michael Khoo is currently a faculty member in the iSchool at Drexel.

[198] Dr. Allan Renear is an Associate Professor in the Graduate School of Library and Information Science at the University of Illinois at Urbana Champaign.

[199] Heer, J et al. 2005. *PREFUSE: A Toolkit for Interactive Information Visualization. ACM CHI 2005*. Portland, Oregon.

[200] Pearce-Moses and Davis, *New Skills,* 2008.

[201] While out of the scope of this dissertation, another outcome of this research is to determine if, which, and how to document and preserve the data sets generated through the text mining process. These include numerical .txt files and images along with all the notes and comparison spreadsheets that I created and used. Another option could be deciding that preserving the records and the software will allow reproducing the results and therefore there is no need to keep the significant amount of matrices produced.

Part IV

[202] Smith, M. 2005. "Eternal Bits." *IEE Spectrum Online*, [cited 17 November 2007]. Available at: http://www.spectrum.iee.org/print/1568

[203] The term is used in reference to storage systems in which access to the records is restricted. It also makes a contrast to the notion of "black box" discussed earlier in this chapter. On this idea see, Caplan, P. 2005. "Building a dark archive in the Sunshine State: A case study." In *Conference Proceedings Archiving 2005*, 9-13. Springfield, Virginia: IS&T.

[204] Upward, *Structuring the Records Continuum*,1998.

[205] See the role of archivists in post-custodial times in the next project: Paradigm Project. *Personal Archives Accessible in Digital Media* [cited 3 April 2008]. Available at: http://www.paradigm.ac.uk/workbook/collection-development/post-custodial.html

[206] All of this is pressing for digital archives of private institutions that due to lack of resources, legal vacuum, or lack of electronic records management directions, are at risk of remaining in limbo during and beyond the records retention period.

[207] The Metadata Timeline in Appendix II shows the evolution of hardware and software used in the institution since 1988 until the present.

[208] Considering that records were used as templates, it can be inferred that only the latest versions of records were used to create the newer version and thus that older records were rarely accessed.

[209] The term "black box" refers to the box in the plains that record what happens in the plain's cockpit and is indestructible. The idea is that black boxes are inaccessible unless an accident occurs and they are recovered to find out what happened after the fact.

[210] Both terms belong to: Pearce-Moses, *A Glossary of Archival Terminology*.

[211] Original order in shared record-keeping environments needs research attention. In Aleph's case, original order is understood as the structure of the shared directory which in turn is organized according to provenance. However, within some virtual folders in which records exist in disarray and groups, series, or sub-series have not been created, the only way in which these can be sorted is by date or alphabetical order. This does not mean that there is other than a provenance relationship between them.

[212] The second InterPARES report on authenticity informed by the case studies concludes that the security of the system in which the electronic records are kept during their active stage is an indicator of their authenticity. See InterPARES, Authenticity Task Force. 2005." Authenticity Task Force Report" [cited 6 October 2006]. Available from the InterPARES website at: http://www.InterPARES.org/display_file.cfm?doc=ip1_atf_report.pdf INTERPARES

[213] American Institute for Conservation (AIC). *Definitions of Conservation Terminology* [cited 5 April 2008]. Available from the AIC website at: http://aic.stanford.edu/geninfo/defin.html

[214] See, Digital Preservation Coalition 2001-2008. "Transfer Procedure and Guidelines," *Preservation Management of Digital Materials: The Handbook.* Originally compiled by Beagrie, N. and Jones M [cited 28 March 2008]. Available from the Digital Preservation Coalition website at: http://www.dpconline.org/graphics/handbook/. Note that the recommendations in the Handbook are issued for records maintained in controlled record-keeping environments.

[215] My original idea was to keep the networked server with all its contents for research purposes to test technical obsolescence. This will not be possible as it was decided to clean the contents and donate the server.

[216] The metadata timeline has limitations. According to the accounting book, in 1994 the foundation also purchased a server. However, the foundation had more than one server. Besides the file server, others were used to provide other services (firewall, development, email, etc.). It is hard to say which server was used for what. Also, the brands and models of the servers are not well detailed in accounting book entries.

[217] Ross, S and Gow, A.1999. *Digital Archaeology Rescuing Neglected and Damaged Data and Damaged Resources* A JISC/NPO Study within the Electronic Libraries (elib) Programme on the Preservation of Electronic Materials, Humanities Advanced Technology and Information Institute HATII, [cited 3 March 2008]. Available at: http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf

[218] Considering that the server will remain unplugged and offline it is logical to think that a RAID 5 configuration is overkill. The purchase was made at a point in which the plan was to maintain the server running.

[219] For a list of the applications installed in the dark archive see Appendix II Metadata Timeline.

[220] Tripwire for Windows is an audit and control software used in high traffic servers with several ports of access. See Tripwire, *Tripwire for Servers/Manager,* [cited 6 April 2008]. Available at: http://www.tripwire.com/products/servers/index.cfm

[221] This is the same technology that is implemented in IR software such as DSpace when files are ingested and throughout records maintenance.

[222] New data storage technologies offer systems that get turned off and on according to needs to avoid wear and to save energy. See Copan Systems, MAID Technology at http://www.copansys.com/architecture/index.shtml

[223] I could not find out what type of algorithm is used by Total Commander to conduct file comparison. In Total Commander's Wiki the entry on file comparison explains that the program compares byte by byte. It also explains that files can be opened in binary form or in text form and that a hex editor can be run. From this I infer that the comparison involves looking at every character in both files making sure that they match to detect any changes or missing parts in the files. See, Total Commander Wiki, [cited 3 March 2008]. Available at: http://www.ghisler.ch/wiki/index.php/Main_Page

[224] These inventories are generated automatically through a command line.

[225] At the time I could not test the financial database, which had access restrictions. Its functionality was tested in 2007 when I closed the digital archiving process in the dark archive.

[226] I used the director's records for all the tests because of all the staff members' records present in the networked server his cover the widest time span.

[227] In a character set such as for example ASCII, control codes do not represent writing symbols but actions such as tabs, escape, printing actions, etc.

228 To understand the cause of the rendering differences I submitted both versions of the files to a Hex editor. The differences in glitches are caused by differences in the control codes and how these are interpreted by the rendering program.

[229] If needed the files can be migrated "on demand" for access using FileMerlin.

[230] The fact that the records and systems had the same structure as the networked server was appreciated by the president's secretary who will, if needed, access the records until the server is sent to the commercial records management company as soon as the foundation is officially closed by the Inspectorate of Justice.

[231] Currently, there is only one possible password-protected access to the dark archive under the name of administrator. The password is under the custody of the foundation's president and his secretary. While officially only they can access the contents of the server, if they give the password to others it will have to be determined who is the "administrator" who changed the files. The issue of trust and password security is faced by all of those who manage any kind of restricted digital repository.

[232] I consulted on many issues with Shane Williams for this dissertation. For example, he suggested the use of Tripwire for Windows as an audit and control software and helped me explore the issues of changes in file properties.

[233] Ball, C. 2005. "Beyond Data About Data: The Litigator's Guide to Metadata," [cited 21 April 2008].Available from http://www.craigball.com/metadata.pdf

[234] In this dissertation I used the last modified date as reference in the different studies because it is the more stable date and time metadata, the closest to the creation date.

[235] Ball, "Beyond Data About Data," 2005.

[236] Galloway, personal communication, 2008.

[237] Holsworth D. 2007. "Installment on Preservation Strategies for Digital Archives. Version 1.0." *Digital Curation Manual,* [cited 4 March 2008]. Available at: http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-strategies/preservation-strategies.pdf

[238] PREMIS Working Group .2005. *Data Dictionary for Preservation Metadata. OCLC, RLG*, [cited 12 March 2008]. Available at: http://www.oclc.org/research/projects/pmwg/premis-final.pdf

[239] Preserving the original archive differs from what InterPARES 2 concluded in 2005. In their preservation report they suggest that preserving the original electronic record is not reasonable or practical. Since representation is ever-changing and vulnerable, there is no such thing as a preserved original electronic record. Preservation can only attempt to reproduce or represent electronic records over time. It is the appropriateness and integrity of the reproduction that is at stake and has to be negotiated and authenticated by the preserver every time. InterPARES does not prescribe a preservation method, but rather a set of administrative processes and sub-processes that function as a decision making tree to assure the adequacy of any kind of preservation strategy. The main assumption under this approach is that technology is not a problem that InterPARES needs to solve; it will be there once archival and preservation decisions have been made See InterPARES 2 Preservation Task Force (nd) "Preservation Task Force Report," [cited 6 March 2008]. Available from InterPARES website at: http://www.InterPARES.org/display_file.cfm?doc=ip1_ptf_report.pdf

[240] McKemmish, S. 1997. "Yesterday, Today and Tomorrow: A Continuum of Responsibility." In *Records Continuum Research Group*, [cited 15 March 2008]. Available from Monash University website at: http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/smckp2.html

Part V

[241]The term domesticating belongs to: Choksy, C. E.B. 2006. *Domesticating Information: Managing documents Inside the Organization.* Maryland: The Scarecrow Press.

[242] Kirschenbaum M.G.2008. *Mechanisms: New Media and the Forensic Imagination*. MIT. In this book, he stresses the stability of electronic media, even of those objects that are inherently fragile or meant by their authors to disappear.

[243] In their study of personal archives of academics Kaye J. et al found that individuals keep the same structure in their analog than in their electronic archive. The difference with the case that I studied is that the analog is the institutional archive (the project files) with a clear numbering system and its structure and content is different than that one of the natural electronic archive. At Aleph, when staff members maintained paper files for their own uses, while the structure or contents were not the same in both systems, how organized or disorganized was the creator that maintains the file could be paralleled in both environments. See Kaye J. et al. 2006. "To Have and to Hold: Exploring the Personal Archive." *CHI'06 Proceeding,* Montreal, 275-284.

244 At the time in which these records were created and used, all of those roles are fulfilled by the paper copies.

245 Upward,. "Structuring the Records Continuum," 1998.

246 Ham, *Understanding Archives and Manuscripts*, 1993.

247 See Helen Tibbo on You Tube talking about her "natural archive." [cited 12 March 2008]. Available at: http://www.youtube.com/watch?v=JbVz6NCADdQ

248 Derrida J. 1995. *Archive Fever: A Freudian Impression*. Chicago: The University of Chicago Press.

249 Dr. Marija Dalbello, Assistant Professor in SCILS at Rutgers University made this comment. I liked the metaphor very much. It ties well with Derrida's idea that archives are impressions that at some point might form a concept and with Borge's Aleph by showing the past in the present.

[250] Taylor, G. 1996. *Cultural selection*. New York: Basic Books.

[251] This relates to Hillary Jenkinson's writing about the naturalness of archives. Jenkinson, *Manual of Archives Administration,* 1937.

[252] Pearce-Moses, *Glossary of Archival Terminology*.

[253] This definition matches Hillary Jenkinson's conception of archives prior to and after being transferred to an archival institution.

[254] Kaplan, E. 2000. "We are What We Collect, We Collect What We Are: Archives and the Construction of Identity." *The American Archivist*, 63, 126-151.

[255] Email communication with Luis Priamo, historic photography scholar from Argentina with whom I "conversed" through email every Sunday during my eight years of studies in Austin.

[256] Jenkinson, *Manual of Archives Administration*, 1937.

[257] It is not a coincidence that in my dissertation committee there are two anthropologists, Dr. Victoria Horwitz and Dr. Patricia Galloway, both practice archaeology and ethnography.

[258] I presented this idea for the first time in May of 2006 at the New Skills for the Digital Era Colloquia in Washington D.C.

[259] After our interview in 2005, and having witnessed the archiving process that had been happening since 2004 Pedro, the executive director told me that if the foundation had continued they would have had to implement some type of electronic record-keeping system.

[260] Leuski, *eArchivarious*, 2003

[261] Explanation of the preservation measures belong to: Cornell University Library, 2003. "Digital Preservation Strategies." *Digital Preservation Management: Implementing short term strategies for long term problems*, [cited 23 March 2008]. Available at: http://www.library.cornell.edu/iris/tutorial/dpm/terminology/strategies.html

[262] This name was given to distinguish the server from the dark archive but it does not have any technical bases to it. It was a way of establishing the difference with the networked server that in this translation I call dark archive to match it with the way in which digital archives with no public access are called.

[263] Cornell, "Digital Preservation Strategies." See Refreshing and Modified Refreshing at: http://www.icpsr.umich.edu/dpm/dpm-eng/terminology/strategies.html

VITA

María Esteva was born in Buenos Aires, Argentina on April 29, 1962, the daughter of Ana Renée Dickstein de Esteva and Hugo Esteva. She has five siblings, Matías, Ana, Maximiliano, Alejo, and Beltrán. After completing her high school education at the Instituto Don Jaime in Bella Vista, Buenos Aires, in 1979 she attended the School of Nutrition at the University of Buenos Aires from which she received a BS in Nutrition in 1983. From 1989 to 1991, she completed internships in rare book conservation at the Library of Congress, at the Folger Shakespeare Library, and at the University of Iowa Center for the Book in the United States. Upon returning to Argentina she worked for ten years coordinating cultural heritage conservation projects in museums, libraries, and archives with the support of Fundación Aleph. In 2001 she obtained a foundation's fellowship to study in the School of Information at the University of Texas at Austin where she graduated with a Masters in Science in Information Studies and an Advanced Certificate in Preservation Administration. In 2003 she started the doctoral program in the same School. Throughout her doctoral studies she obtained the Marietta Daniels Presidential Scholarship. She has a son named Simón González Esteva

Permanent Address: San Martín 1039, (1661) Bella Vista, Provincia de Buenos Aires, Argentina

This dissertation was typed by the author.